INFORMATION MEASURES FOR SOURCES WITH MEMORY AND THEIR APPLICATION TO HYPOTHESIS TESTING AND SOURCE CODING

by

Ziad Rached

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
August 2002

Abstract

In this work, we investigate Shannon's and Rényi's information measure rates for finite-alphabet time-invariant Markov sources of arbitrary order and arbitrary initial distributions, along with their application to hypothesis testing and source coding. We also study, using information-spectrum techniques, Csiszár's forward and reverse cutoff rates for the hypothesis testing problem between general sources with memory (including all non-ergodic or non-stationary sources) with arbitrary alphabet (countable or uncountable).

We first provide a computable expression for the Kullback-Leibler divergence rate, $\lim_{n\to\infty}\frac{1}{n}D(p^{(n)}\|q^{(n)})$, between two Markov sources described by the probability distributions $p^{(n)}$ and $q^{(n)}$, respectively. We illustrate it numerically and examine its rate of convergence. Similarly, we provide a formula for the Shannon entropy rate, $\lim_{n\to\infty}\frac{1}{n}H(p^{(n)})$, of Markov sources and examine its rate of convergence. As an application to hypothesis testing, we provide an alternative simple proof for Stein's Lemma for testing between stationary irreducible Markov sources.

We also address the existence and the computation of the Rényi α-divergence rate, $\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$, between Markov sources, where $\alpha > 0$ and $\alpha \neq 1$. We provide numerical examples and examine its rate of convergence. We also investigate the limits of the Rényi divergence rate as $\alpha \to 1$ and as $\alpha \downarrow 0$. Similarly, we provide a formula for the Rényi entropy rate, $\lim_{n\to\infty}\frac{1}{n}H_\alpha(p^{(n)})$, of Markov sources. We also study its rate of convergence and its limits as $\alpha \to 1$ and as $\alpha \downarrow 0$. As an application to source coding, we present a generalization of Campbell's variable-length source coding theorem for discrete memoryless sources to Markov sources. This provides a new operational characterization for the Rényi entropy rate. The main tools used to obtain Shannon's and Rényi's information measure rates results are the theory of non-negative matrices and Perron-Frobenius theory.

We next establish an operational characterization for the Rényi α-divergence rate, by showing, using an information-spectrum approach, that the Csiszár forward β-cutoff rate for the hypothesis testing problem between general sources with memory is given by the $\liminf$ α-divergence rate with $\alpha = \frac{1}{1-\beta}$. The Csiszár forward β-cutoff rate ($\beta < 0$) for hypothesis testing is defined as the largest rate $R_0 \geq 0$ such that for all rates $0 < E < R_0$, the best (i.e., smallest) probability of type 1 error of sample size-n tests with probability of type 2 error $\leq e^{-nE}$ is asymptotically vanishing as $e^{-n\beta(E-R_0)}$. We also demonstrate that, under some conditions on the large deviation spectrum, the Csiszár reverse β-cutoff rate for the general hypothesis testing problem is given by the $\limsup$ α-divergence rate with $\alpha = \frac{1}{1-\beta}$. The Csiszár reverse β-cutoff rate ($\beta > 0$) for hypothesis testing is defined as the smallest rate $R_0 \geq 0$ such that for all rates $0 < R_0 < E$, the best (i.e., largest) correct probability of type 1 of sample size-n tests with probability of type 2 error $\leq e^{-nE}$ is asymptotically vanishing as $e^{-n\beta(E-R_0)}$. Furthermore, we investigate the important classes of discrete memoryless sources and sources that satisfy the hypotheses of the Gärtner-Ellis Theorem for which the forward and reverse β-cutoff rates are computable. Finally, we conclude with observations and remarks along with several possible directions for future work.

Acknowledgments

I would like to thank my supervisor, Dr. Fady Alajaji for his support, his continuous encouragement and his helpful suggestions and comments throughout the completion of this work.

I would like to also thank Dr. Lorne Campbell who luminously pointed out the importance of Perron-Frobenius theory for this thesis and Dr. Po-Ning Chen for fruitful discussions about the hypothesis testing cutoff rate problem.

Finally, I would like to gratefully acknowledge the financial support received from the Natural Sciences and Engineering Research Council (NSERC), the Ontario Graduate Scholarship (OGS), the Ontario Graduate Scholarship in Science and Technology (OGSST), the Ontario Premier Research Excellence Award (PREA), the Department of Mathematics and Statistics, and the School of Graduate Studies.

Contents

1 Introduction 1

1.1 Literature Review 1

1.2 Contributions 4

1.3 Thesis Overview 5

2 Preliminaries: Non-Negative Matrices and Discrete Markov Sources 8

2.1 Non-Negative Matrices and Perron-Frobenius Theory 8

2.2 Discrete Markov Sources and Stochastic Matrices 21

3 Shannon's Information Measure Rates for Finite-Alphabet Markov Sources 33

3.1 Kullback-Leibler Divergence Rate 35

3.1.1 First-Order Markov Sources 35

3.1.2 $\boldsymbol{k}$-th Order Markov Sources 39

3.2 Shannon Entropy Rate 40

3.3 Numerical Examples 42

3.4 Hypothesis Testing Error Exponent For Stationary Irreducible Markov Sources 48

4 Rényi's Information Measure Rates for Finite-Alphabet Markov Sources 56

4.1 Rényi Divergence Rate 59

4.1.1 First-order Markov Sources 59

4.1.2 $\boldsymbol{k}$-th Order Markov Sources 71

4.1.3 Numerical Examples 72

4.2 Interchangeability of Limits 77

4.2.1 Limit as $\boldsymbol{\alpha \to 1}$ 77

4.2.2 Limit as $\boldsymbol{\alpha \downarrow 0}$ 84

4.3 Rényi's Entropy Rate 85

4.4 A Variable-Length Source Coding Theorem 91

5 Csiszár's Forward Cutoff Rate for Hypothesis Testing Between General Sources with Memory 96

5.1 Preliminaries 97

5.2 Hypothesis Testing Forward $\boldsymbol{\beta}$-Cutoff Rate 100

5.3 Numerical Examples 116

5.4 Properties of $\boldsymbol{\tau}$ and $\boldsymbol{\sigma(R)}$ 124

6 Csiszár's Reverse Cutoff Rate for Hypothesis Testing Between General Sources with Memory 128

6.1 Preliminaries and Problem Formulation 129

6.2 Hypothesis Testing Reverse $\boldsymbol{\beta}$-Cutoff Rate 134

6.3 Properties of $\boldsymbol{\lambda}$ 156

7 Conclusion and Future Work 158

7.1 Summary 158

7.2 Future Work 159

Bibliography 161

Vita 169

Chapter 1

Introduction

The first subject of this thesis is the investigation of Shannon's and Rényi's information measure rates for finite-alphabet time-invariant Markov sources, along with their application to hypothesis testing and source coding. The second subject is the investigation of Csiszár's cutoff rates for the hypothesis testing problem between general sources with memory (not necessarily Markovian, stationary, ergodic, etc.). In this chapter, we present the literature review of articles upon which our research is based. We then specify the main contributions of the thesis and present its outline.

1.1 Literature Review

The concept of entropy as a measure of information of a random variable was first introduced by Shannon in his celebrated 1948 paper [55]. He investigated the properties of entropy and its applications to source coding in the context of discrete memoryless sources (DMS). Since then, a considerable amount of research has focused on providing new measures of information and extending Shannon's results for more general sources (Markov, stationary, ergodic, etc.). A particular alternative measure to Shannon's entropy that brought the attention of many researchers is the Rényi entropy [52], $H_\alpha(p)$, or entropy of order α. An operational characterization of Rényi's entropy in the context of source coding was first given by Campbell in [13]. He showed that, for DMS, Rényi's entropy plays a role analogous to the Shannon entropy in variable-length source coding when the cost function in the coding problem is exponential as opposed to linear. This occurs in many applications where the processing cost of decoding is high or the buffer overflow due to long codewords is important. From this work, a natural question arises: how can one generalize Shannon's and Campbell's variable-length source coding theorems for DMS to more general sources with memory, such as Markov sources. This led us to investigate Shannon's entropy rate, $\lim_{n\to\infty} \frac{1}{n} H(p^{(n)})$, and Rényi's entropy rate, $\lim_{n\to\infty} \frac{1}{n} H_\alpha(p^{(n)})$, for Markov sources. Previous work on the computation of Shannon's entropy rate for stationary and irreducible Markov sources may be found in [10], [18], [25]. In [25], the author showed the existence of the Shannon entropy rate for arbitrary Markov sources (not necessarily stationary, irreducible, etc.), but he did not provide the computational details.

The Rényi entropy and the Rényi entropy rate have revealed several operational characterizations in the problem of fixed-length source coding [14, 20], variable-length source coding [11, 34], error exponent calculations [23], and other areas [1, 6, 8, 46].

Other important measures, primarily introduced in the hypothesis testing problem between DMS, are the Kullback-Leibler divergence [40], $D(p\|q)$ and the Rényi divergence [52], $D_\alpha(p\|q)$, or the α-divergence. The application of the Kullback-Leibler divergence can be found in many areas such as approximation of probability distributions [17], [38], signal processing [36], [37], [22], pattern recognition [9], [16], etc. In [26], Gray proved that the Kullback-Leibler divergence rate, $\lim_{n\to\infty}\frac{1}{n}D(p^{(n)}\|q^{(n)})$, exists between a stationary source $p^{(n)}$ and a Markov source $q^{(n)}$. This result can also be found in [59, p. 27]. In [42], the authors noted that the Kullback-Leibler divergence rate between ergodic Markov sources exits. Also, in [56], Shields presented two examples for non-Markovian sources for which the Kullback-Leibler divergence rate does not exist.

The Rényi divergence rate, $\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$, has played a significant role in certain hypothesis testing questions [39, 44, 45]. In [44], [45], the author evaluated the Rényi divergence rate between two Markov sources under the restriction that the initial probabilities are strictly positive.

The β-cutoff rate concept, for source coding and hypothesis testing, was first introduced in [20] for DMS. In [14], the authors generalized the source coding β-cutoff rate for DMS to general sources (not necessarily stationary, ergodic, etc.) using an *information spectrum* philosophy which was developed by Han and Verdú [27]. With the aid of this method, Verdú and Han obtained a general formula for the capacity of arbitrary single-user channels (not necessarily information stable, stationary, etc.) without feedback [58]. In [30], Han addressed at length many information theoretic problems using the information spectrum approach which is a very powerful tool that applies to general sources (not necessarily Markovian, stationary, ergodic, etc.) and general alphabets (countable or uncountable). Several results from this book were recently published in the IEEE Transactions on Information Theory. In particular, Han investigated in [28] the optimal exponent problem for the probability of decoding error and correct decoding in fixed-length source coding. In [29], he studied the hypothesis testing problem between general sources with memory. Specifically, he examined the optimal exponent problem for the type 2 probability of testing error, as well as the type 2 probability of correct testing subject to an exponential error constraint on the type 1 probability of testing error.

1.2 Contributions

The contributions of this thesis (parts of which appeared in [3], [4], [47]–[51]) are as follows:

- Computable expressions for the Kullback-Leibler divergence rate and for the Shannon entropy rate for arbitrary finite-alphabet Markov sources along with their rate of convergence.

- Computable expressions for the Rényi α-divergence rate and for the Rényi entropy rate for arbitrary finite-alphabet Markov sources along with their rate of convergence.

- Sufficient conditions under which the Rényi information measure rates for Markov sources reduce to the Shannon information measure rates as $\alpha \to 1$ and the interchangeability of limits between n and α as $n \to \infty$ and as $\alpha \downarrow 0$.

- Generalization of Campbell's variable-length source coding theorem for DMS to Markov sources which provides an operational characterization for the Rényi entropy rate.

- A simple proof of Stein's Lemma for hypothesis testing between stationary irreducible Markov sources.

- A generalization of Csiszár's forward and reverse β-cutoff rates for hypothesis testing between DMS to general sources with memory of arbitrary alphabet. This yields an operational characterization for the α-divergence rate. An examination of the important classes of DMS and Markov sources for which the forward and reverse β-cutoff rates are computable is also provided.

1.3 Thesis Overview

The thesis is organized in the following manner.

In Chapter 2, we present some useful properties and results from linear algebra, specifically the theory of non-negative matrices and Perron-Frobenius theory. We also present some useful properties and results for discrete stochastic processes, specifically discrete Markov chains.

In Chapter 3, we provide a computable expression for the Kullback-Leibler divergence rate between time-invariant Markov sources with finite alphabet and arbitrary initial distributions. The result is first proved for first-order Markov sources, and is then extended for Markov sources of arbitrary order. We illustrate it numerically and examine its rate of convergence. Similarly, we address the computation and the rate of convergence for the Shannon entropy rate of Markov sources. Using the formula for the Kullback-Leibler divergence rate, we provide a simple alternative proof of Stein's Lemma for testing between stationary irreducible Markov sources.

In Chapter 4, we generalize Nemetz's result by establishing a formula for the α-divergence rate between two time-invariant Markov sources with arbitrary initial distributions and illustrate it numerically. The result is first proved for first-order Markov sources, and is then extended for Markov sources of arbitrary order. We then show that if the probability transition matrix P associated with the Markov source under $p^{(n)}$ is absolutely continuous with respect to the probability transition matrix Q associated with the Markov source under $q^{(n)}$ and if the initial distribution p under $p^{(n)}$ is absolutely continuous with respect to the initial distribution q under $q^{(n)}$, then the Rényi divergence rate reduces to the Kullback-Leibler divergence rate as $\alpha \to 1$. We also show that the interchangeability of limits as $n \to \infty$ and as $\alpha \downarrow 0$ is always valid. Furthermore, we address similar questions for the Rényi entropy rate. As an application to source coding, we provide a new operational characterization for the Rényi entropy rate by generalizing Campbell's variable-length source coding theorem for DMS to Markov sources.

In Chapter 5, we review relevant previous results by Han on the optimal asymptotic exponent of the probability of testing error. We then derive a general expression for the forward β-cutoff rate for hypothesis testing between arbitrary sources. We demonstrate that the liminf α-divergence rate, where $\alpha = \frac{1}{1-\beta}$ and $\beta < 0$, provide the expression for the forward β-cutoff rate. We also provide numerical examples based on DMS using Cramer's Theorem [12].

In Chapter 6, we review relevant previous definitions and results by Csiszár and Han on the optimal asymptotic exponent of the probability of correct testing. Under two conditions on the log likelihood ratio large deviation spectrum, $\rho(R)$, we show that the reverse β-cutoff rate is given by the $\limsup$ α-divergence rate, where $\alpha = \frac{1}{1-\beta}$ and $0 < \beta < \beta_{\max}$, where $\beta_{\max}$ is the largest $\beta < 1$ for which the $\limsup$ $\frac{1}{1-\beta}$-divergence rate is finite. For $\beta_{\max} \leq \beta < 1$, we provide an upper bound on the reverse cutoff rate. In particular, we examine finite-alphabet independent and identically distributed (i.i.d.) observations and sources that satisfy the hypotheses of the Gärtner-Ellis Theorem [12]. We show that in these cases, the conditions on $\rho(R)$ are satisfied and that the reverse cutoff rate admits a simple form. We also provide several numerical examples to illustrate our results. The main tools used in obtaining the forward and reverse cutoff rates results are large deviation theory and the information spectrum approach.

In Chapter 7, we conclude with a summary along with several directions for future work.

Chapter 2

Preliminaries: Non-Negative Matrices and Discrete Markov Sources

2.1 Non-Negative Matrices and Perron-Frobenius Theory

We begin with some useful definitions and important properties about determinants that can be found in any text book in linear algebra such as [32]. Throughout, $A := (a_{ij})$ denotes an $M \times M$ square matrix.

Definition 2.1 A pair of numbers j_k and j_p in a permutation $(j_1, j_2, \ldots, j_M)$ form an *inversion* if $j_k > j_p$ while $k < p$, that is, if a larger number in the permutation precedes a smaller one. Each permutation $j = (j_1, j_2, \ldots, j_M)$ has a certain number of inversions associated with it, denoted briefly by $t(j)$. The permutation is called *odd* or *even* according to whether the number $t(j)$ is odd or even.

Definition 2.2 The *determinant* of A, denoted by $\det(A)$ or $|A|$, is defined as

$$|A| = \sum_j (-1)^{t(j)} a_{1j_1} a_{2j_2} \cdots a_{Mj_M}, \tag{2.1}$$

where j varies over all the $M!$ permutations of $1, 2, \ldots, M$.

Lemma 2.1 If B is obtained from A by multiplying one of its rows (or columns) by a scalar k, then $|B| = k|A|$.

Lemma 2.2 If B is obtained by interchanging two rows (or columns) of A, then $|B| = -|A|$.

Lemma 2.3 If B is obtained from A by adding the elements of its i-th row (or column) to the corresponding elements of its j-th row (or column) multiplied by a scalar α, then $|B| = |A|$.

Lemma 2.4 Suppose that the entries of A are functions of some parameter α. Let $|A|_i$ be the determinant obtained from A by replacing the elements in the i-th row by their derivatives with respect to α and leaving the other rows unchanged. Then

$$|A|' = \sum_{i=1}^{M} |A|_i,$$

where $|A|'$ is the derivative of $|A|$ with respect to α.

Proof: If we differentiate (2.1), we get that

$$|A|' = \sum_i (-1)^{t(j)} (a_{1j_1} a_{2j_2} \dots a_{Mj_M})',$$

where j varies over all $M!$ permutations of $1, 2, \dots, M$. By the product rule of derivatives

$$(a_{1j_1} a_{2j_2} \dots a_{Mj_M})' = a'_{1j_1} a_{2j_2} \dots a_{Mj_M} + a_{1j_1} a'_{2j_2} \dots a_{Mj_M} + \dots + a_{1j_1} a_{2j_2} \dots a'_{Mj_M}.$$

Therefore

$$\begin{aligned} |A|' &= \sum_j (-1)^{t(j)} a'_{1j_1} a_{2j_2} \cdots a_{Mj_M} + \sum_j (-1)^{t(j)} a_{1j_1} a'_{2j_2} \cdots a_{Mj_M} \\ &\quad + \dots + \sum_j (-1)^{t(j)} a_{1j_1} a_{2j_2} \cdots a'_{Mj_M}. \end{aligned}$$

Hence, we conclude that $|A|' = \sum_i |A|_i$. □

Definition 2.3 A *minor* of order $M-1$ of A is defined to be the determinant of a submatrix of A obtained by deleting one row and one column. The minor obtained by deleting the i-th row and the j-th column is denoted by L_{ij}, $(1 \le i, j \le M)$. The *cofactor* A_{ij} of an element a_{ij} is given by: $A_{ij} = (-1)^{i+j} L_{ij}$.

Lemma 2.5 The determinant of A can be computed as follows:

$$|A| = a_{i1} A_{i1} + a_{i2} A_{i2} + \dots + a_{iM} A_{iM},$$

or similarly,

$$|A| = a_{1j} A_{1j} + a_{2j} A_{2j} + \dots + a_{Mj} A_{Mj}.$$

Definition 2.4 A *right eigenvector*, b, corresponding to an *eigenvalue* λ, is a nonzero vector such that $Ab = \lambda b$. A *left eigenvector*, a, corresponding to λ, is a nonzero vector such that $aA = \lambda a$. Note that a is a row vector while b is a column vector.

Definition 2.5 A *Jordan block* $J_s(\lambda)$ corresponding to an eigenvalue λ of A is a $s \times s$ upper triangular matrix of the form

$$J_s(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \lambda & 1 \\ 0 & \dots & \dots & \dots & \lambda \end{bmatrix}.$$

Definition 2.6 An $M \times M$ *Jordan matrix* J for A is of the form

$$J = \begin{bmatrix} J_{n_1}(\lambda_1) & 0 & \dots & \dots & 0 \\ 0 & J_{n_2}(\lambda_2) & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & J_{n_r}(\lambda_r) \end{bmatrix}, \quad n_1 + n_2 + \ldots + n_r = M,$$

where 0 denotes a zero matrix (i.e., all entries are zeros) with appropriate dimension.

Theorem 2.1 [32, p. 126] Let λ_i, $i = 1, \ldots, r$ be the eigenvalues of A (not necessarily distinct). There is an invertible matrix S such that

$$A = SJS^{-1}.$$

The following limiting behavior result of A can be proved using its Jordan form.

Theorem 2.2 [32, p. 138] The matrix A^m converges to the zero matrix 0 as $m \to \infty$ iff the eigenvalues of A have modulus strictly less than 1.

Lemma 2.6 If all the eigenvalues of A have modulus strictly less than 1, then $I - A$ is invertible.

Proof: Note first that if λ is an eigenvalue of A, then $1 - \lambda$ is an eigenvalue of $I - A$. Indeed, if $Ab = \lambda b$, then

$$(A - I)b = Ab - Ib = \lambda b - b = (\lambda - 1)b.$$

Therefore, all the eigenvalues of $I - A$ are non-zero. Hence, it is invertible since its determinant is non-zero (the determinant is equal to the product of the eigenvalues by simply considering the Jordan block form of A). □

Definition 2.7 The *algebraic multiplicity* of an eigenvalue λ is its multiplicity as a root of the characteristic equation $\det(A - \lambda I) = 0$, where I is the identity matrix.

Let us also recall some definitions and results about non-negative matrices and Perron Frobenius theory. Most of what follows may be found in [54, Chapter 1], [24, Chapter 4], and [32, Chapter 8].

Definition 2.8 A Matrix or a vector is *positive* if all its components are positive and *non-negative* if all its components are non-negative.

Throughout, unless otherwise stated, A denotes an $M \times M$ non-negative matrix ($A \geq 0$) with elements a_{ij}. The ij-th element of A^m is denoted by $a_{ij}^{(m)}$. We write $i \to j$ if $a_{ij}^{(m)} > 0$ for some positive integer m, and we write $i \not\to j$ if $a_{ij}^{(m)} = 0$ for every positive integer m.

Definition 2.9 Two indices i and j *communicate* ($i \leftrightarrow j$) if $i \to j$ and $j \to i$.

Definition 2.10 If $i \to j$ but $j \not\to i$ for some index j, then the index i is called *inessential*. An index which leads to no index at all (this arises when A has a row of zeros) is also called inessential.

Definition 2.11 An index i is essential if $i \to j$ implies $i \leftrightarrow j$, and there is at least one j such that $i \to j$.

With these definitions, it is possible to partition the set of indices $\{1, 2, \ldots, M\}$ into disjoint sets, called *classes*. All essential indices (if any) can be subdivided into *essential classes* in such a way that all the indices belonging to one class communicate, but cannot lead to an index outside the class. Moreover, all inessential indices (if any) may be divided into two types of *inessential classes*: *self-communicating* classes and *non self-communicating* classes. Each self-communicating inessential class contains inessential indices which communicate with each other. A non self-communicating inessential class is a singleton set whose element is an index which does not communicate with any index (including itself).

Definition 2.12 A matrix is *irreducible* if its indices form a single essential class; i.e., if every index communicates with every other index.

Definition 2.13 The *period* of an index i, denoted $d(i)$, is defined as the greatest common divisor (gcd) of those values of n for which $a_{ii}^{(n)} > 0$. If the period is 1, the index is *aperiodic*, and if the period is 2 or more, the index is *periodic.*

Proposition 2.1 [54, p. 17] In a communicating class, all indices have the same period.

Definition 2.14 An irreducible matrix is said to be *periodic* with period d, if the period of any one (and so of each one) of its indices satisfies $d > 1$, and is said to be *aperiodic* if $d = 1$.

Proposition 2.2 [54, p. 15] By renumbering the indices (i.e., by performing row and column permutations), it is possible to put a non-negative matrix A in the *canonical form*

$$A = \begin{bmatrix} A_1 & \ldots & 0 & 0 & \ldots & 0 & \ldots & \ldots & 0 \\ 0 & \ldots & 0 & 0 & \ldots & 0 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & A_h & 0 & \ldots & 0 & \ldots & \ldots & 0 \\ A_{h+11} & \ldots & A_{h+1h} & A_{h+1} & \ldots & 0 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ A_{g1} & \ldots & A_{gh} & A_{gh+1} & \ldots & A_g & \ldots & \ldots & 0 \\ A_{g+11} & \ldots & A_{g+1h} & A_{g+1h+1} & \ldots & A_{g+1g} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ A_{l1} & \ldots & A_{lh} & A_{lh+1} & \ldots & A_{lg} & A_{lg+1} & \ldots & 0 \end{bmatrix}$$

where A_i, $i = 1, \ldots, g$, are irreducible square matrices (periodic in general), and in each row $i = h+1, \ldots, g$ at least one of the matrices $A_{i1}, A_{i2}, \ldots, A_{ii-1}$ is not zero. The matrix A_i for $i = 1, \ldots, h$ corresponds to the essential class C_i; while the matrix A_i for $i = h+1, \ldots, g$ corresponds to the self-communicating inessential class C_i. The other diagonal block sub-matrices which correspond to non self-communicating classes C_i, $i = g+1, \ldots, l$, are 1×1 zero matrices. In every row $i = g+1, \ldots, l$ any of the matrices $A_{i1}, \ldots, A_{ii-1}$ may be zero.

Definition 2.15 A class C_j is *reachable* from another class C_i where $j = 1, \ldots, l$ and $i = h+1, \ldots, l$ if $A_{ij} \neq 0$, or if for some $i_1, \ldots, i_c$, $A_{ii_1} \neq 0$,$A_{i_1 i_2} \neq 0$,...,$A_{i_c,j} \neq 0$, where c is at most $l-1$ (since there are l classes).

Remark: c can be viewed as the number of steps needed to reach class C_j starting from class C_i. Note that from the canonical form of A, the class C_j is reachable from class C_i if $A_{ij}^{(c)} \neq 0$ for some $c = 1, \ldots, l-1$, where $A_{ij}^{(c)}$ is the ij-th submatrix of A^c. Note also that no class can be reached from any of the classes $C_1, \ldots, C_h$ since they are essential classes.

Example: Consider the following non-negative matrix A along with its canonical form A_c.

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad A_c = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

The canonical form A_c is obtained by permuting the first and third rows and columns and the second and sixth rows and columns of A. Note that A_c has 2 essential classes, $C_1 = \{1, 2\}$ and $C_2 = \{3, 4\}$, 1 inessential self-communicating class, $C_3 = \{5, 6\}$, and 1 inessential non self-communicating class, $C_4 = \{7\}$. Also, note that the class C_1 is not reachable from the class C_2 (since C_1 and C_2 are essential classes), however it is reachable from C_3 and C_4.

Proposition 2.3 (Perron) [24, p. 115] If A is positive, then A has a real positive eigenvalue λ with algebraic multiplicity 1 that is greater than the magnitude of each other eigenvalue. There is a positive left (right) eigenvector, a (b), corresponding to λ, where a is a row vector and b is a column vector.

The theory of non-negative matrices was initiated by Perron for positive matrices and generalized later by Frobenius for irreducible matrices. The key idea is that if A is irreducible, then $(I + A)^{M-1} > 0$, where I is the identity matrix. The latter inequality follows directly from the definition of an irreducible matrix. Indeed, if A is irreducible, then for all $i, j = 1, \ldots, M$, $a_{ij}^{(n)} > 0$, for some $1 \leq n \leq M - 1$.

Proposition 2.4 (Frobenius) [24, p. 115] If A is irreducible, then A has a real positive eigenvalue λ that is greater than or equal to the magnitude of each other eigenvalue. There is a positive left (right) eigenvector, a (b), corresponding to λ, where a is a row vector and b is a column vector.

The proof relies on the fact that $(I+A)^{M-1} > 0$ and the fact that if λ is an eigenvalue of A, then $1 + \lambda$ is an eigenvalue of $I + A$. Also, $I + A$ and A have exactly the same eigenvectors.

Proposition 2.5 [32, p. 492] Suppose A is irreducible and let R_i, $i = 1, \ldots, M$ denote the sum of the i-th row. Also, let $R_{\max} = \max\{R_1, \ldots, R_M\}$ and $R_{\min} = \min\{R_1, \ldots, R_M\}$. Then the largest positive real eigenvalue λ satisfies

$$R_{\min} \leq \lambda \leq R_{\max}.$$

The following lemma follows by appropriately modifying the proof of the above proposition.

Lemma 2.7 If A is irreducible and the row sums are not all identical, then the largest positive real eigenvalue λ satisfies,

$$R_{\min} < \lambda < R_{\max}.$$

Proof: Let λ be the largest positive real eigenvalue of A with associated strictly positive left eigenvector a, which exists by Proposition 2.4. Without loss of generality a can be normalized, i.e., the sum of its components is equal to 1. Let $\mathbf{1}^t$ be the row vector

$$\mathbf{1}^t = (1, \ldots, 1).$$

Note that $a\mathbf{1} = 1$, where t denotes the transpose operation. We have $aA = \lambda a$. Hence $aA\mathbf{1} = \lambda a\mathbf{1} = \lambda$. On the other hand

$$\begin{aligned} aA\mathbf{1} &= a(R_1, \ldots, R_M)^t \\ &< a(R_{\max}, \ldots, R_{\max})^t \\ &= \sum_{i=1}^{M} a_i R_{\max} \\ &= R_{\max} \end{aligned}$$

Therefore $\lambda < R_{\max}$. Similarly, we can show that $\lambda > R_{\min}$. Finally we conclude that

$$R_{\min} < \lambda < R_{\max}.$$

□

Proposition 2.6 Suppose A is irreducible. Let λ be the largest positive real eigenvalue with associated right positive eigenvector b. Then $A^m \leq \lambda^m C$ (i.e., $a_{ij}^{(m)} \leq \lambda^m c_{ij}$), for all $m = 1, 2, \ldots$, where $C = (\frac{\max_{1\leq k\leq M} b_k}{\min_{1\leq k\leq M} b_k})$ is a matrix with identical entries that are independent of m.

Proof: If $Ab = \lambda b$, then $A^m b = \lambda^m b$. We have that

$$\begin{aligned}
\lambda^m(\max_{1\leq k\leq M} b_k) &\geq \lambda^m b_i \\
&= \sum_{j=1}^{M} a_{ij}^{(m)} b_j \\
&\geq (\min_{1\leq k\leq M} b_k)\sum_{j=1}^{M} a_{ij}^{(m)} \\
&\geq (\min_{1\leq k\leq M} b_k) a_{ij}^{(m)},
\end{aligned}$$

for all $i = 1, \ldots, M$ and $j = 1, \ldots, M$. Since $b > 0$, we obtain the desired result.

□

Proposition 2.7 [32, p. 508] If A is irreducible, then the largest positive real eigenvalue has algebraic multiplicity 1.

Proof: Let $B = A/\lambda$, where λ is the largest positive real eigenvalue of A. By the previous corollary, B^m is bounded above by C for all $m = 1, 2, \ldots$ Note that the largest positive real eigenvalue of B is 1. The block corresponding to this eigenvalue in the *Jordan canonical form* of B must have size 1×1, because otherwise, the entries of this block diverge as $m \to \infty$ which contradicts the fact that B^m is uniformly bounded for all $m = 1, 2, \ldots$ □

Proposition 2.8 [41, p. 371] The eigenvalues of a matrix are continuous functions of the entries of the matrix.

This proposition follows from that fact that the roots of a polynomial are continuous functions of its coefficients, and the fact that the eigenvalues are the roots of the characteristic equation of the matrix.

Proposition 2.9 [32, p. 372] Let $A(t)$ be an $M \times M$ matrix whose entries are all differentiable functions at $t = 0$. Assume that λ is an eigenvalue of $A(0) = A$ of algebraic multiplicity 1, and that $\lambda(t)$ is an eigenvalue of $A(t)$, for small t, such that $\lambda(0) = \lambda$. Let a (b) be the left (right) eigenvector corresponding to λ, such that $ab = 1$. Then

$$\lambda'(t)|_{t=0} = aA'(t)|_{t=0}b.$$

Proof: By the previous proposition, for all sufficiently small t there is an eigenvalue $\lambda(t)$ of $A(t)$ such that $\lambda(0) = \lambda$. There is also a left (right) eigenvector $a(t)$ ($b(t)$) corresponding to $\lambda(t)$ such that $a(t)b(t) = 1$. If we differentiate this last normalization condition, we obtain the identity

$$a'(t)b(t) + a(t)b'(t) = 0. \tag{2.2}$$

Since $A(t)b(t) = \lambda(t)b(t)$ for all small t, we also have the identity $a(t)A(t)b(t) = \lambda(t)a(t)b(t) = \lambda(t)$. If we differentiate this identity, we obtain

$$\lambda'(t) = a'(t)A(t)b(t) + a(t)A'(t)b(t) + a(t)A(t)b'(t).$$

But since $A(t)b(t) = \lambda(t)b(t)$ and $a(t)A(t) = \lambda(t)a(t)$, we obtain via (2.2) that

$$\lambda'(t) = \lambda(t)\{a'(t)b(t) + a(t)b'(t)\} + a(t)A'(t)b(t) = a(t)A'(t)b(t).$$

Thus

$$\lambda'(t)|_{t=0} = aA'(t)|_{t=0}b.$$

□

2.2 Discrete Markov Sources and Stochastic Matrices

Most of the following can be found in [18, Chapter 4] and [24, Chapter 4].

Definition 2.16 A discrete stochastic process $\{X_1, X_2, \ldots\}$ with finite-alphabet $\mathcal{X} = \{1, 2, \ldots, M\}$ is said to be a *Markov source* of order k if, for $n > k$,

$$Pr\{X_n = i_n | X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots, X_1 = i_1\} =$$
$$Pr\{X_n = i_n | X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots, X_{n-k} = i_{n-k}\},$$

for all $i_1, \ldots, i_n \in \mathcal{X}$.

Define $\{W_n\}$ as the process obtained by *k-step blocking* the Markov source $\{X_n\}$; i.e.,

$$W_n \stackrel{\triangle}{=} (X_n, X_{n+1}, \ldots, X_{n+k-1}).$$

Then

$$Pr\{W_n = w_n | W_{n-1} = w_{n-1}, \ldots, W_1 = w_1\} = Pr\{W_n = w_n | W_{n-1} = w_{n-1}\},$$

and hence, $\{W_n\}$ is a first order Markov source with M^k states. We herein consider Markov sources of first order unless otherwise stated.

Definition 2.17 A stochastic process is said to be *stationary* if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in time index, i.e.,

$$Pr\{X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n\} = Pr\{X_{1+l} = i_1, X_{2+l} = i_2, \ldots, X_{n+l} = i_n\},$$

for every time shift l and for all $i_1, \ldots, i_n \in \mathcal{X}$.

Definition 2.18 A Markov source is said to be *time-invariant* if the conditional probability does not depend on n, i.e., for $n > 1$,

$$Pr\{X_n = j | X_{n-1} = i\} = Pr\{X_2 = j | X_1 = i\}, \quad \text{for all} \quad i, j \in \mathcal{X}.$$

If $\{X_1, X_2, \ldots\}$ is a Markov source, then X_n is called the *state* at time n. A time-invariant Markov source is characterized by its initial state and a *probability transition matrix* $P = (p_{ij})$, $i, j \in \mathcal{X}$, where $p_{ij} = Pr\{X_{n+1} = j | X_n = i\}$. From now on, we will only deal with time-invariant Markov sources.

Definition 2.19 A distribution on the states such that the distribution at time $n+1$ is the same as the distribution at time n is called a *stationary distribution* and is denoted by $\pi = (\pi_1, \ldots, \pi_M)$.

Remark: For a finite-alphabet Markov source with probability transition matrix P, its stationary distribution π always exists [24, p. 110] and can be obtained by solving $\pi P = \pi$. Furthermore, the source is *stationary* if the distribution of its initial state is given by π.

Definition 2.20 A Markov chain is *irreducible* if its probability transition matrix P is irreducible. It is *ergodic* if P is irreducible and aperiodic.

Definition 2.21 The *entropy rate* of a stochastic process $\{X_1, X_2, \ldots\}$ is defined by

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

when the limit exists.

Definition 2.22 We can also define a related quantity for entropy rate:

$$H'(\mathcal{X}) = \lim_{n \to \infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1),$$

when the limit exists.

The two above quantities correspond to two different notions of entropy rate. The first is the per symbol of the n random variables, and the second is the conditional entropy of the last random variable given the past.

Proposition 2.10 [18, p. 64] For a stationary source, $H(\mathcal{X})$ and $H'(\mathcal{X})$ exist and are equal.

Proposition 2.11 [18, p. 66], [25, p. 68] Let $\{X_1, X_2, \ldots\}$ be a Markov source with stationary distribution π and transition matrix P. Then the entropy rate is given by

$$H(\mathcal{X}) = H(X_2|X_1) = -\sum_{i,j\in\mathcal{X}} \pi_i p_{ij} \log p_{ij},$$

if the source is stationary. The same result also holds for irreducible (not necessarily stationary) Markov sources.

Example: *Finite-memory Polya contagion process*: Consider the following source $\{X_1, X_2, \ldots\}$ which is generated according to the following urn scheme as described in [2]: An urn initially contains T balls–R red and S black ($T = R + S$). At the j-th draw, j=1,2,..., we select a ball from the urn and replace it with $1 + \triangle$ balls of the same color ($\triangle > 0$); then, k draws later–after the $(j + k)$-th draw–we retrieve from the urn $\triangle$ balls of the color picked at time j. Let $\rho = R/T < 1/2$, $\sigma = 1 - \rho = S/T$ and $\delta = \triangle/T$. Then, the source $\{X_i\}$ corresponds to the outcomes of the draws from the urn, where

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th ball drawn is red} \\ 0, & \text{if the } i\text{-th ball drawn is black} \end{cases}$$

It was shown in [2] that $\{X_1, X_2, \ldots\}$ is a stationary ergodic Markov source of order k with entropy rate given by

$$H(\mathcal{X}) = H(X_{k+1}|X_k, \ldots, X_1) = \sum_{i=0}^{k} \begin{pmatrix} k \\ i \end{pmatrix} L_i h_b\left(\frac{\rho + i\delta}{1 + k\delta}\right),$$

where

$$L_i = \frac{\prod_{j=0}^{i-1}(\rho + j\delta) \prod_{l=0}^{k-i-1}(\sigma + l\delta)}{\prod_{m=1}^{k-1}(1 + m\delta)},$$

and

$$h_b(a) := -a \log_2 a - (1-a)\log_2(1-a)$$

is the binary entropy function.

Proposition 2.12 [26, p. 40] The *Kullback-Leibler divergence rate* between a stationary source $p^{(n)}$, with stationary distribution π, and a Markov source $q^{(n)}$, with transition matrix $Q = (q_{ij})$, is given by

$$\lim_{n\to\infty} \frac{1}{n} D(p^{(n)} \| q^{(n)}) = -H_p(\mathcal{X}) - \sum_{i,j\in\mathcal{X}} \pi_i p_{ij} \log q_{ij},$$

where $H_p(\mathcal{X})$ is the entropy rate of the stationary source $p^{(n)}$ which exists by Proposition 2.10.

Let us recall some useful results from Perron-Frobenius theory in the context of stochastic matrices. An immediate consequence of Propositions 2.4 and 2.5 is the following result.

Corollary 2.1 Let P be the probability transition matrix for an irreducible Markov source. Then $\lambda = 1$ is an eigenvalue of P which is greater than or equal to the magnitude of each other eigenvalue.

Proposition 2.13 [32, p. 524] Let P be the probability transition matrix for an irreducible Markov source. Also, let a (b) be the left (right) eigenvector associated with the largest positive real eigenvalue $\lambda = 1$ such that $ab = 1$. Also, let $L = ba$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P^i = L.$$

Moreover, there exists a finite positive constant $C = C(P)$ such that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} P^i - L \right\|_\infty \leq \frac{C}{n},$$

for all $n = 1, 2, \ldots$ and $\| \cdot \|_\infty$ is the l_∞ *norm*, where the l_∞ norm of an $M \times M$ matrix A is defined by $\|A\|_\infty \overset{\triangle}{=} \max_{1 \leq i,j \leq M} |a_{ij}|$.

Proof: We have that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} P^i &= \frac{1}{n} \sum_{i=1}^{n} [(P - L)^i + L] & (2.3) \\ &= L + \frac{1}{n} \sum_{i=1}^{n} (P - L)^i \\ &= L + \frac{1}{n} (P - L)(I - (P - L)^n)(I - (P - L))^{-1} & (2.4) \\ &= L + \frac{1}{n} (P - L)(I - P^n + L)(I - (P - L))^{-1}, & (2.5) \end{aligned}$$

where (2.3) follows from the identity $(P - L)^m = P^m - L$ for all $m = 1, 2, \ldots$ (which can be shown by induction on m) and (2.4) follows from the fact that if B is a square matrix such that $I - B$ is invertible, then $\sum_{i=1}^{n} B^i = B(I - B^n)(I - B)^{-1}$. It can be shown that the matrix $I - (P - L)$ is indeed invertible. The equality (2.5) follows also from the identity $(P - L)^m = P^m - L$. The only part in (2.5) that depends on n is the factor $1/n$ and the term P^n. But, by Proposition 2.6, P^n is uniformly bounded as $n \to \infty$. Thus, $\frac{1}{n} \sum_{i=1}^{n} P^i$ converges to L, and the order of convergence is $1/n$.

□

Remark: The left eigenvector a is the unique stationary distribution π of P associated with the largest positive real eigenvalue $\lambda = 1$ and $b^t = (1, \ldots, 1)$.

With the aid of the above proposition and Proposition 2.2, it can be shown that for an arbitrary stochastic matrix P the Cesáro limit, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} P^i$, exists and is computable.

Proposition 2.14 [19, p. 129] Let P be the probability transition matrix for an arbitrary Markov source with associated canonical form as in Proposition 2.2. Let a_i (b_i) be the left (right) eigenvector of P_i associated with $\lambda = 1$ such that $a_i b_i = 1$, for $i = 1, \ldots, h$. Let

$$A = \begin{bmatrix} P_1 & \ldots & 0 \\ 0 & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & P_h \end{bmatrix}, \quad B = \begin{bmatrix} P_{h+11} & \ldots & P_{h+1h} \\ \ldots & \ldots & \ldots \\ P_{g1} & \ldots & P_{gh} \\ P_{g+11} & \ldots & P_{g+1h} \\ \ldots & \ldots & \ldots \\ P_{l1} & \ldots & P_{lh} \end{bmatrix}.$$

Also, let

$$C = \begin{bmatrix} P_{h+1} & \ldots & 0 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ P_{gh+1} & \ldots & P_g & \ldots & \ldots & 0 \\ P_{g+1h+1} & \ldots & P_{g+1g} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ P_{lh+1} & \ldots & P_{lg} & P_{lg+1} & \ldots & 0 \end{bmatrix}, \quad D = \begin{bmatrix} b_1 a_1 & \ldots & 0 \\ 0 & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & b_h a_h \end{bmatrix}.$$

We have the following:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}P^i = \begin{bmatrix} D & 0 \\ (I-C)^{-1}BD & 0 \end{bmatrix},$$

where I is the identity matrix.

Proof: We have that

$$P^n = \begin{bmatrix} A^n & 0 \\ B^{(n)} & C^n \end{bmatrix}.$$

Note that

$$B^{(n)} = BA^{n-1} + CB^{(n-1)},$$

by simply equating the entries of the matrix P^n with the entries of the matrix PP^{n-1}. Therefore

$$\sum_{i=1}^{n} B^{(i)} = B\sum_{i=1}^{n} A^{i-1} + C\sum_{i=1}^{n} B^{(i-1)},$$

where $B^{(0)} := 0$ and $A^0 := I$. Hence

$$\frac{1}{n}\sum_{i=1}^{n} B^{(i)} = B\frac{1}{n}\sum_{i=1}^{n} A^{i-1} + C\frac{1}{n}\sum_{i=1}^{n} B^{(i-1)}. \tag{2.6}$$

By Proposition 2.13

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} P_i^j = b_i a_i,$$

for $i = 1, \ldots, h$, where b_i (a_i) is the right (left) eigenvector corresponding to 1 which is the largest positive real eigenvalue corresponding to all the stochastic matrices P_i such that $a_i b_i = 1$. It follows that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} A^j = D = \begin{bmatrix} b_1 a_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & b_h a_h \end{bmatrix}.$$

If P_i, $i = h+1, \ldots, g$ has all row sums identical, then by Proposition 2.5, its largest positive real eigenvalue is less than 1. Otherwise, by Lemma 2.7, its largest positive real eigenvalue is less than 1. Hence, all the eigenvalues of C have modulus less than 1. Therefore C^n converges to the zero matrix 0, and hence

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} C^j = 0.$$

Letting $n \to \infty$ in (2.6), we conclude that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} B^{(j)} = (I - C)^{-1} BD,$$

where $I - C$ is invertible by Lemma 2.6, and hence the desired result. □

Proposition 2.15 (Perron's formula) [53, Section 5] Let $\lambda_0, \lambda_1, \ldots, \lambda_r$ be the eigenvalues of A, with algebraic multiplicities $m_0, m_1, \ldots, m_r$, respectively. Define $\psi_t(\lambda)$ by

$$A(\lambda) = |\lambda I - A| = (\lambda - \lambda_t)^{m_t} \psi_t(\lambda), \quad t = 0, \ldots, r,$$

such that $\psi_t(\lambda)$ are polynomials of degree $M - m_t$ which differ from zero for $\lambda = \lambda_t$. Then, we have identically for all $i, j = 1, \ldots, M$ and $k = 1, 2, 3, \ldots$

$$a_{ij}^{(k)} = \sum_{t=0}^{r} \frac{1}{(m_t - 1)!} D_\lambda^{m_t - 1} \left[\frac{\lambda^k A_{ij}(\lambda)}{\psi_t(\lambda)} \right]_{\lambda = \lambda_t},$$

where $A_{ij}(\lambda)$ is the cofactor of the ij-th element of $\lambda I - A$. In this equation, $D_\lambda^{m_t-1}$ denotes the derivative of order $m_t - 1$ with respect to λ, evaluated at $\lambda = \lambda_t$.

Note that Perron's formula permits to express an arbitrary element $a_{ij}^{(k)}$ of the matrix A^k in terms of the eigenvalues of A and the cofactors of the matrix $\lambda I - A$.

Lemma 2.8 [53, p. 10] Let $A(\lambda) = |\lambda I - A|$. Denote by $A_{ij}(\lambda)$ the cofactor of the ij-th element of the matrix $\lambda I - A$. I is the $M \times M$ identity matrix. Then

$$\frac{dA(\lambda)}{d\lambda} = \sum_{i=1}^{M} A_{ii}(\lambda).$$

Proof: By applying Lemma 2.4 to the determinant $A(\lambda)$, the i-th row of $A_i(\lambda)$ consists of zeros except the i-th position which is 1. By Lemma 2.5, expanding each $A_i(\lambda)$ along this row yields the desired result. □

Lemma 2.9 [53, p. 10] Suppose in addition to the previous lemma that $\lambda = 1$ and each row of A sums to 1. Then

$$A_{i1}(1) = A_{i2}(1) = \cdots = A_{iM}(1),$$

for all $i = 1, 2, \ldots, M$.

Proof: This statement follows by using the properties of determinants in Lemma 2.1, Lemma 2.2, and Lemma 2.3. □

Proposition 2.16 [53, p. 17] Let P be the probability transition matrix for an ergodic Markov source. Then the stationary distribution π is given by

$$\pi_i = \frac{P_{ii}(1)}{\sum_j P_{jj}(1)}, \quad i = 1, \ldots, M,$$

where $P_{ij}(1)$ denotes the cofactor of the ij-th entry of the matrix $I - P$, and I is the identity matrix.

Proof: Applying Proposition 2.15 to P yields

$$p_{ij}^{(k)} = \frac{1}{(m_0-1)!} D_\lambda^{m_0-1} \left[\frac{\lambda^k P_{ij}(\lambda)}{p_0(\lambda)}\right]_{\lambda=1} + \sum_{t=1}^{r} \frac{1}{(m_t-1)!} D_\lambda^{m_t-1} \left[\frac{\lambda^k P_{ij}(\lambda)}{p_t(\lambda)}\right]_{\lambda=\lambda_t}, \quad (2.7)$$

in which $\lambda_0 = 1, \lambda_1, \ldots, \lambda_r$ are the eigenvalues of P and $m_0, m_1, \ldots, m_r$ their respective multiplicities, so that $m_0 + m_1 + \cdots + m_r = M$. The polynomials $p_0(\lambda), p_1(\lambda), \ldots,$ $p_r(\lambda)$ are defined by

$$P(\lambda) = (\lambda-1)^{m_0} p_0(\lambda) = (\lambda-\lambda_t)^{m_t} p_t(\lambda), \quad t = 1, \ldots, r,$$

where

$$p_0(1) \neq 0, \quad p_t(\lambda_t) \neq 0, \qquad t = 1, \ldots, r.$$

This relationship has a particular importance for the ergodic Markov chain associated with P since $\lambda_0 = 1$ is a simple eigenvalue, i.e., $m_0 = 1$. In this case, (2.7) assumes the form

$$p_{ij}^{(k)} = \frac{P_{ij}(1)}{p_0(1)} + \sum_{t=1}^{r} \frac{1}{(m_t-1)!} D_\lambda^{m_t-1} \left[\frac{\lambda^k P_{ij}(\lambda)}{p_t(\lambda)}\right]_{\lambda=\lambda_t}. \quad (2.8)$$

By Lemma 2.9, $P_{ij}(1) = P_{ii}(1)$. Also, since $P(\lambda) = (\lambda-1)p_0(\lambda)$, then, $P'(\lambda) = p_0(\lambda) + (\lambda-1)p_0'(\lambda)$, and, $P'(1) = p_0(1) \neq 0$.

But by Lemma 2.8 $P'(\lambda) = \sum_i P_{ii}(\lambda)$. Therefore, $P'(1) = \sum_i P_{ii}(1) \neq 0$.

For simplicity let

$$\frac{1}{(m_t-1)!} D_\lambda^{m_t-1} \left[\frac{\lambda^k P_{ij}(\lambda)}{\lambda_t^k p_t(\lambda)}\right]_{\lambda=\lambda_t} \triangleq Q_{ijt}(k).$$

Clearly, $Q_{ijt}(k)$ represents a polynomial in k of degree not greater than $(m_t - 1)$, and we can therefore write

$$Q_{ijt}(k) = \sum_{h=0}^{m_t-1} Q_{ijt}^{(h)} k^h,$$

where the $Q_{ijt}^{(h)}$ represent some specific numbers which do not depend on k. We conclude that (2.8) can be written as

$$p_{ij}^{(k)} = p_i + \sum_{t=1}^{r} Q_{ijt}(k)\lambda_t^k,$$

where

$$p_i = \frac{P_{ii}(1)}{P'(1)} = \frac{P_{ii}(1)}{\sum_j P_{jj}(1)}.$$

The magnitude of all the remaining eigenvalues of P are less than unity. Since $Q_{ijt}(k)$ are polynomials of finite degree in k, it follows that

$$\lim_{k\to\infty} p_{ij}^{(k)} = p_i, \quad i = 1, 2, \ldots, M,$$

since

$$\lim_{k\to\infty} k^h \lambda^k = 0.$$

To show the above equality, it is sufficient to prove that

$$\lim_{k\to\infty} k^h |\lambda|^k = 0. \tag{2.9}$$

We have the following two cases: if $|\lambda| = 0$ then (2.9) is obvious. Otherwise, $0 < |\lambda| < 1$. In this case,

$$\begin{aligned} \lim_{k\to\infty} \log k^h |\lambda|^k &= \lim_{k\to\infty} (h \log k + k \log |\lambda|) \\ &= \lim_{k\to\infty} k \left(h \frac{\log k}{k} + \log |\lambda| \right) \\ &= -\infty, \end{aligned}$$

since $\lim_{k\to\infty} \frac{\log k}{k} = 0$ by l'Hôpital's rule and $\log |\lambda| < 0$. Therefore, (2.9) also holds in this case.

□

Chapter 3

Shannon's Information Measure Rates for Finite-Alphabet Markov Sources

Let $\{X_1, X_2, \ldots\}$ be a first-order time-invariant Markov source with finite-alphabet $\mathcal{X} = \{1, \ldots, M\}$. Consider the following two different probability laws for this source. Under the first law,

$$Pr\{X_1 = i\} =: p_i \quad \text{and} \quad Pr\{X_{k+1} = j | X_k = i\} =: p_{ij}, \quad i, j \in \mathcal{X},$$

so that

$$p^{(n)}(i^n) := Pr\{X_1 = i_1, \ldots, X_n = i_n\} = p_{i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}, \quad i_1, \ldots, i_n \in \mathcal{X},$$

while under the second law the initial probabilities are q_i, the transition probabilities are q_{ij}, and the n-tuple probabilities are $q^{(n)}$. Let $p = (p_1, \ldots, p_M)$ and

$q = (q_1, \ldots, q_M)$ denote the initial distributions under $p^{(n)}$ and $q^{(n)}$ respectively.

The Kullback-Leibler divergence [40] between two distributions $\hat{p}$ and $\hat{q}$ defined on $\mathcal{X}$ is given by

$$D(\hat{p}\|\hat{q}) = \sum_{i \in \mathcal{X}} \hat{p}_i \log \frac{\hat{p}_i}{\hat{q}_i},$$

where the base of the logarithm is arbitrary. One natural direction for further studies is the investigation of the Kullback-Leibler divergence rate

$$\lim_{n\to\infty} \frac{1}{n} D(p^{(n)}\|q^{(n)})$$

between two probability distributions $p^{(n)}$ and $q^{(n)}$ defined on $\mathcal{X}^n$, where

$$D(p^{(n)}\|q^{(n)}) = \sum_{i^n \in \mathcal{X}^n} p^{(n)}(i^n) \log \frac{p^{(n)}(i^n)}{q^{(n)}(i^n)},$$

for sources with memory. In [26], Gray proved that the Kullback-Leibler divergence rate exists between a stationary source $p^{(n)}$ and a time-invariant Markov source $q^{(n)}$ (Proposition 2.12). This result can also be found in [59, p. 27]. To the best of our knowledge, this is the only result available in the literature about the existence and the computation of the Kullback-Leibler divergence rate between sources with memory. In the sequel, we provide a computable expression for the Kullback-Leibler divergence rate between two arbitrary time-invariant finite alphabet Markov sources. This expression, which is proved in a straightforward manner using results from the theory of non-negative matrices and Perron-Frobenius theory, has a readily usable form, making it appealing for various analytical studies and applications involving the divergence between systems with memory.

3.1 Kullback-Leibler Divergence Rate

3.1.1 First-Order Markov Sources

We first assume that the time-invariant Markov source $\{X_1, X_2, \ldots\}$ is of order one. Later, we generalize the results for sources of arbitrary order k. Let p and q be the initial distributions with respect to $p^{(n)}$ and $q^{(n)}$ respectively. Also, let P and Q be the probability transition matrices with respect to $p^{(n)}$ and $q^{(n)}$ respectively. Without loss of generality, we may assume that p and P are absolutely continuous with respect to q and Q respectively (i.e., $q_i = 0 \Rightarrow p_i = 0$ and $q_{ij} = 0 \Rightarrow p_{ij} = 0$, for all $i, j \in \mathcal{X}$), because otherwise the Kullback-Leibler divergence rate is infinite. We have the following results.

Theorem 3.1 Suppose that the Markov source $\{X_1, X_2, \ldots\}$ is irreducible under $p^{(n)}$ and $q^{(n)}$. Let

$$S(X_2|X_1 = i) \triangleq \sum_{j \in \mathcal{X}} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Then, the Kullback-Leibler divergence rate between $p^{(n)}$ and $q^{(n)}$ is given by

$$\lim_{n \to \infty} \frac{1}{n} D(p^{(n)} \| q^{(n)}) = \sum_{i \in \mathcal{X}} \pi_i S(X_2|X_1 = i),$$

where $\pi = (\pi_1, \ldots, \pi_M)$ is the unique stationary distribution of P.

Proof: We have that

$$\begin{aligned}
&\frac{1}{n}D(p^{(n)}\|q^{(n)}) = \\
&\quad \frac{1}{n}\sum_{i\in\mathcal{X}}[p(X_1=i)+\cdots+p(X_{n-1}=i)]S(X_2|X_1=i)+ \\
&\quad \frac{1}{n}\sum_{i\in\mathcal{X}}p(X_1=i)\log\frac{p(X_1=i)}{q(X_1=i)},
\end{aligned}$$

which can be also written as

$$\begin{aligned}
\frac{1}{n}D(p^{(n)}\|q^{(n)}) &= \frac{1}{n}p(I+P+\cdots+P^{n-2})V && (3.1)\\
&\quad +\frac{1}{n}\sum_{i\in\mathcal{X}}p_i\log\frac{p_i}{q_i}, && (3.2)
\end{aligned}$$

where

$$V^t = (S(X_2|X_1=1),\ldots,S(X_2|X_1=M)).$$

Note that (3.2) approaches 0 as $n\to\infty$. Hence, by Proposition 2.13, we obtain that

$$\lim_{n\to\infty}\frac{1}{n}p(I+P+\cdots+P^{n-2})V = pLV,$$

where

$$\begin{aligned}
L &= ba = (1,\ldots,1)^t(\pi_1,\ldots,\pi_M)\\
&= \begin{bmatrix} \pi_1 & \pi_2 & \ldots & \pi_M \\ \pi_1 & \pi_2 & \ldots & \pi_M \\ \vdots & \vdots & \vdots & \vdots \\ \pi_1 & \pi_2 & \ldots & \pi_M \end{bmatrix}.
\end{aligned}$$

Thus

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}D(p^{(n)}\|q^{(n)}) &= p\begin{bmatrix}\pi_1 & \pi_2 & \dots & \pi_M\\ \pi_1 & \pi_2 & \dots & \pi_M\\ \vdots & \vdots & \vdots & \vdots\\ \pi_1 & \pi_2 & \dots & \pi_M\end{bmatrix}V\\ &= \sum_{i\in\mathcal{X}}\pi_i S(X_2|X_1=i)\end{aligned}$$

□

Theorem 3.2 Suppose that the Markov source $\{X_1, X_2, \ldots\}$ under $p^{(n)}$ and $q^{(n)}$ is arbitrary[1] (not necessarily irreducible, stationary, etc.). Let the canonical form of P be as in Proposition 2.2. Also, let B, D and C be as defined in Proposition 2.14. Then, the Kullback-Leibler divergence rate between $p^{(n)}$ and $q^{(n)}$ is given by

$$\lim_{n\to\infty}\frac{1}{n}D(p^{(n)}\|q^{(n)}) = p\begin{bmatrix} D & 0\\ (I-C)^{-1}BD & 0\end{bmatrix}V,$$

where

$$V^t = (S(X_2|X_1=1), \ldots, S(X_2|X_1=M)),$$

and I is the identity matrix with same dimensions as the matrix C.

[1] Since p and P are absolutely continuous with respect to q and Q respectively, it follows that $p^{(n)}$ is absolutely continuous with respect to $q^{(n)}$. Hence, some restriction on their behavior is induced. For instance, if P is irreducible, Q must be irreducible. However, it is possible to have Q irreducible and P reducible. So, in general, Q and P do not necessarily have the same number of classes.

Proof: As in the previous theorem, we have that

$$\frac{1}{n}D(p^{(n)}\|q^{(n)}) = \frac{1}{n}p(I+P+\cdots+P^{n-2})V \tag{3.3}$$

$$+\frac{1}{n}\sum_{i\in\mathcal{X}} p_i \log\frac{p_i}{q_i}. \tag{3.4}$$

Then, the desired result follows immediately from Proposition 2.14.

□

Theorem 3.3 The rate of convergence of the Kullback-Leibler divergence rate between arbitrary $p^{(n)}$ and $q^{(n)}$ is of the order $1/n$.

Proof: Clearly, the rate of convergence of (3.4) to 0 is of the order $1/n$. In Proposition 2.13, it is proved that the rate of convergence of the Cesáro sum of an irreducible stochastic matrix is of the order $1/n$. On the other hand, if P is not irreducible, let P_i, $i = 1,\ldots,h$, be the sub-matrices corresponding to essential classes and let P_i, $i = h+1,\ldots,g$ be the sub-matrices corresponding to inessential classes as in Proposition 2.2. For $i = 1,\ldots,h$, each P_i is stochastic and irreducible; so its Cesáro-sum is of the order $1/n$ by Proposition 2.13. Now, for $i = h+1,\ldots,g$, every P_i is irreducible and hence, by Proposition 2.6, we have that

$$P_i^n \leq \lambda_i^n G_i, \quad i = h+1,\ldots,g, \tag{3.5}$$

where λ_i is the largest positive real eigenvalue of P_i, and G_i is a matrix with identical entries that are independent of n. Therefore

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n} P_i^j &\leq \frac{1}{n}\sum_{j=1}^{n}\lambda_i^j G_i \\ &= \frac{1}{n}\frac{\lambda_i(1-\lambda_i^n)}{1-\lambda_i}G_i,\end{aligned}$$

for $i = h+1, \ldots, g$. If P_i has all row sums identical then $\lambda_i < 1$ by Proposition 2.5 and the fact that P is stochastic. Otherwise, $\lambda_i < 1$ by Lemma 2.7. Hence, the Cesáro sum of P_i, $i = h+1, \ldots, g$ is of the order $1/n$. By considering the Cesáro sum of the canonical form of P, we get that the rate of convergence of (3.3) is of the order $1/n$. Therefore the rate of convergence of the Kullback-Leibler divergence rate is of the order $1/n$. □

3.1.2 *k*-th Order Markov Sources

Now, suppose that the Markov source has an arbitrary order k. Define $\{W_n\}$ as the process obtained by k-step blocking the Markov source $\{X_n\}$; i.e.,

$$W_n := (X_n, X_{n+1}, \ldots, X_{n+k-1}).$$

Then $\{W_n\}$ is a first order Markov source with M^k states. Let $p_{w_{n-1}w_n} := Pr(W_n = w_n | W_{n-1} = w_{n-1})$. Let $p = (p_1, \ldots, p_{M^k})$ and $q = (q_1, \ldots, q_{M^k})$ denote the arbitrary initial distributions of W_1 under $p^{(n)}$ and $q^{(n)}$ respectively. Also, let p_{ij} and q_{ij} denote the transition probability that W_n goes from index i to index j under $p^{(n)}$ and $q^{(n)}$ respectively, $i, j = 1, \ldots, M^k$. Then clearly $D(p^{(n)} \| q^{(n)})$ can be written as

$$\begin{aligned} \frac{1}{n} D(p^{(n)} \| q^{(n)}) &= \frac{1}{n} p(I + P + \cdots + P^{n-2})V \\ &\quad + \frac{1}{n} \sum_{i \in \mathcal{X}^k} p(W_1 = i) \log \frac{p(W_1 = i)}{q(W_1 = i)}, \end{aligned}$$

where

$$V^t = (S(W_2|W_1 = 1), \ldots, S(W_2|W_1 = M^k)).$$

It follows directly that Theorems 3.2 and 3.3 also hold for a Markov source of arbitrary order k.

3.2 Shannon Entropy Rate

The existence and the computation of the Shannon entropy rate of an arbitrary time-invariant finite-alphabet Markov source can be directly deduced from the existence and the computation of the Kullback-Leibler divergence rate. Indeed, if $q^{(n)}$ is stationary memoryless with uniform marginal distribution, then

$$D(p^{(n)}\|q^{(n)}) = n\log M - H(p^{(n)}).$$

Therefore

$$\lim_{n\to\infty}\frac{1}{n}D(p^{(n)}\|q^{(n)}) = \log M - \lim_{n\to\infty}\frac{1}{n}H(p^{(n)}). \tag{3.6}$$

We have the following corollaries.

Corollary 3.1 Suppose that the Markov source $\{X_1, X_2, \ldots\}$ under $p^{(n)}$ is irreducible. Let

$$H(X_2|X_1 = i) \overset{\triangle}{=} -\sum_{j\in\mathcal{X}} p_{ij}\log p_{ij}.$$

Then, the Shannon entropy rate of $p^{(n)}$ is given by

$$\lim_{n\to\infty}\frac{1}{n}H(p^{(n)}) = \sum_{i\in\mathcal{X}} \pi_i H(X_2|X_1 = i),$$

where $\pi = (\pi_1, \ldots, \pi_M)$ is the unique stationary distribution of P.

Proof: Obtained directly by plugging $q_{ij} = 1/M$ in Theorem 3.1 and using (3.6).

□

Corollary 3.2 Let the canonical form of P be as in Proposition 2.2. Also, let B, D and C be as defined in Proposition 2.14. Then, the Shannon entropy rate is given by

$$\lim_{n\to\infty}\frac{1}{n}H(p^{(n)}) = p\begin{bmatrix} D & 0 \\ (I-C)^{-1}BD & 0 \end{bmatrix} V,$$

where

$$V^t = (H(X_2|X_1=1),\ldots,H(X_2|X_1=M)),$$

and I is the identity matrix with same dimensions as the matrix C.

Proof: Note that P^i, $i=1,2,\ldots$ is a stochastic matrix[2]. Hence,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}(I+P+\cdots+P^{n-2})\mathbf{1}^t &= \lim_{n\to\infty}\frac{n-1}{n}\mathbf{1}^t \\ &= \mathbf{1}^t\end{aligned}$$

which yields that

$$\lim_{n\to\infty}\frac{1}{n}(I+P+\cdots+P^{n-2})$$

is a stochastic matrix. Therefore,

$$\begin{bmatrix} D & 0 \\ (I-C)^{-1}BD & 0 \end{bmatrix}$$

is also a stochastic matrix. Hence,

$$p\begin{bmatrix} D & 0 \\ (I-C)^{-1}BD & 0 \end{bmatrix}\begin{bmatrix}\log M \\ \vdots \\ \log M\end{bmatrix} = p\begin{bmatrix}\log M \\ \vdots \\ \log M\end{bmatrix} = \log M.$$

[2] We have that $P\mathbf{1}^t = \mathbf{1}^t$, where $\mathbf{1} = (1,\ldots,1)$ and t is the transpose operation. Using this fact and the fact that $P^i = PP^{i-1}$, the result follows by mathematical induction on i.

Then, the corollary follows directly by plugging $q_{ij} = \frac{1}{M}$ in Theorem 3.2 and using (3.6).

□

Corollary 3.3 The rate of convergence of the Shannon entropy rate of $p^{(n)}$ is of the order $1/n$.

3.3 Numerical Examples

In this section, we use the natural logarithm for simplicity.

Example 1: Let P and Q be two possible probability transition matrices for a first order Markov source $\{X_1, X_2, \ldots\}$ (not stationary and not irreducible) defined as follows:

$$P = \begin{bmatrix} 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 4/7 & 2/7 & 1/7 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 2/3 & 0 \\ 1/4 & 0 & 0 & 3/4 & 0 & 0 & 0 \\ 2/5 & 2/5 & 0 & 0 & 1/5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 1/2 & 0 & 1/4 & 0 & 0 \end{bmatrix},$$

and

$$Q = \begin{bmatrix} 1/3 & 0 & 0 & 2/3 & 0 & 0 & 0 \\ 0 & 0 & 2/7 & 1/7 & 4/7 & 0 & 0 \\ 0 & 0 & 1/5 & 0 & 0 & 4/5 & 0 \\ 1/6 & 0 & 0 & 5/6 & 0 & 0 & 0 \\ 1/5 & 2/5 & 0 & 0 & 2/5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 1/4 & 0 & 1/2 & 0 & 0 \end{bmatrix}.$$

Let $p = (3/7, 0, 1/7, 0, 1/7, 2/7, 0)$ and $q = (2/8, 0, 3/8, 0,\ 1/8, 2/8, 0)$ be two possible initial distributions under $p^{(n)}$ and $q^{(n)}$, respectively. In canonical form, P and Q can be rewritten as

$$P = \begin{bmatrix} 1/3 & 2/3 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 3/4 & 0 & 0 & 0 \\ 0 & 0 & 2/5 & 0 & 1/5 & 2/5 & 0 \\ 4/7 & 0 & 0 & 2/7 & 1/7 & 0 & 0 \\ 1/2 & 0 & 1/4 & 0 & 1/4 & 0 & 0 \end{bmatrix},$$

and

$$Q = \begin{bmatrix} 1/5 & 4/5 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 2/3 & 0 & 0 & 0 \\ 0 & 0 & 1/6 & 5/6 & 0 & 0 & 0 \\ 0 & 0 & 1/5 & 0 & 2/5 & 2/5 & 0 \\ 2/7 & 0 & 0 & 1/7 & 4/7 & 0 & 0 \\ 1/4 & 0 & 1/4 & 0 & 1/2 & 0 & 0 \end{bmatrix},$$

simply by permuting the first and third rows and columns and the second and sixth rows and columns. Note that P has 2 essential classes, 1 inessential self-communicating class and 1 inessential non self-communicating class. Accordingly, the initial distributions are rewritten as $p = (1/7, 2/7, 3/7, 0, 1/7, 0, 0)$ and $q = (3/8, 2/8, 2/8, 0, 1/8, 0, 0)$, after permuting the first and third indices and the second and sixth indices. We obtain the following.

n	$\frac{1}{n}D(p^{(n)}\|q^{(n)})$
10	0.05323
50	0.03626
100	0.03415

By Theorem 3.2, the Kullback-Leibler divergence rate is equal to 0.032. Clearly, as n gets large $\frac{1}{n}D(p^{(n)}\|q^{(n)})$ is closer to the Kullback-Leibler divergence rate. We also obtain the following.

n	$\frac{1}{n}H(p^{(n)})$
10	0.54366
50	0.50877
100	0.50442

By Corollary 3.2, the Shannon entropy rate is equal to 0.50008. Clearly, as n gets large $\frac{1}{n}H(p^{(n)})$ is closer to the Shannon entropy rate.

Example 2: Consider the Markov source $\{X_i\}$ of order 2 generated according a variation of the Polya urn scheme as described in the example of Chapter 3. The process $\{W_n\}$ such that each random variable W_n is a 2-step blocking of $\{Z_n\}$, i.e.

$$W_n = (Z_n, Z_{n+1}),$$

is a first order stationary ergodic Markov source with 4 states. The probability transition matrix P of $\{W_n\}$ is given by

$$P = \begin{bmatrix} \frac{\sigma+2\delta}{1+2\delta} & \frac{\rho}{1+2\delta} & 0 & 0 \\ 0 & 0 & \frac{\sigma+\delta}{1+2\delta} & \frac{\rho+\delta}{1+2\delta} \\ \frac{\sigma+\delta}{1+2\delta} & \frac{\rho+\delta}{1+2\delta} & 0 & 0 \\ 0 & 0 & \frac{\sigma}{1+2\delta} & \frac{\rho+2\delta}{1+2\delta} \end{bmatrix},$$

where $\rho + \sigma = 1$. Suppose that the urn contains initially 3 red balls and 5 black balls. Denote by $p^{(n)}$ the joint distribution of the source and P its transition matrix if $\triangle = 1$. Denote by $q^{(n)}$ the joint distribution of the source and Q its transition matrix if $\triangle = 2$. In this case

$$P = \begin{bmatrix} 7/10 & 3/10 & 0 & 0 \\ 0 & 0 & 6/10 & 4/10 \\ 6/10 & 4/10 & 0 & 0 \\ 0 & 0 & 5/10 & 5/10 \end{bmatrix}, \quad Q = \begin{bmatrix} 9/12 & 3/12 & 0 & 0 \\ 0 & 0 & 7/12 & 5/12 \\ 7/12 & 5/12 & 0 & 0 \\ 0 & 0 & 5/12 & 7/12 \end{bmatrix}.$$

The initial distributions under $p^{(n)}$ and $q^{(n)}$ are respectively $p = (30/72, 15/72, 15/72, 12/72)$ and $q = (35/80, 15/80, 15/80, 15/80)$. We obtain the following.

n	$\frac{1}{n}D(p^{(n)}\|q^{(n)})$
10	0.0046
50	0.00512
100	0.00519

By Theorem 3.1, the Kullback-Leibler divergence rate is equal to 0.005254. Clearly, as n gets large $\frac{1}{n}D(p^{(n)}\|q^{(n)})$ is closer to the Kullback-Leibler divergence rate. We also obtain the following.

n	$\frac{1}{n}H(p^{(n)})$
10	0.3887
50	0.5981
100	0.6243

By Corollary 3.1, the Shannon entropy rate is equal to 0.6505. Clearly, as n gets large $\frac{1}{n}H(p^{(n)})$ is closer to the Shannon entropy rate.

Example 3: Suppose that the Markov source is of order 2 under $p^{(n)}$ and $q^{(n)}$ respectively. Let $\{W_1, W_2, \ldots\}$ be the process obtained by 2-step blocking the Markov source. Let P and Q be two possible transition matrices for $\{W_1, W_2, \ldots\}$ defined as follows:

$$P = \begin{bmatrix} 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2/5 & 3/5 & 0 & 0 \\ 0 & 0 & 1/6 & 5/6 \end{bmatrix},$$

and

$$Q = \begin{bmatrix} 3/4 & 1/4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 7/8 & 1/8 & 0 & 0 \\ 0 & 0 & 2/3 & 1/3 \end{bmatrix}.$$

Let $p = (1/8, 3/8, 2/8, 2/8)$ and $q = (1/7, 2/7, 3/7, 1/7)$ denote two possible initial distributions of W_1 under $p^{(n)}$ and $q^{(n)}$ respectively. The set of indices $\{1, 2, 3\}$ forms an essential class, while the singleton set $\{4\}$ forms a self-communicating non-essential class. Hence, P and Q are not irreducible. Note also that both $p^{(n)}$ and $q^{(n)}$ are not stationary. We obtain the following.

n	$\frac{1}{n}D(p^{(n)}\|q^{(n)})$
10	0.2982
50	0.3253
100	0.3277

By Theorem 3.2, the Kullback-Leibler divergence rate is equal to .3301. Clearly, as n gets large $\frac{1}{n}D(p^{(n)}\|q^{(n)})$ is closer to the Kullback-Leibler divergence rate. We also obtain the following.

n	$\frac{1}{n}H(p^{(n)})$
10	0.4618
50	0.4175
100	0.4116

By Corollary 3.2, the Shannon entropy rate is equal to 0.4057. Clearly, as n gets large $\frac{1}{n}H(p^{(n)})$ is closer to the Shannon entropy rate.

3.4 Hypothesis Testing Error Exponent For Stationary Irreducible Markov Sources

Let us first recall the binary hypothesis testing problem. Consider a sequence of random variables $\{X_1, \ldots, X_n\}$ which is generated according to some distribution $p^{(n)}$ under the null hypothesis H_1 and generated according to some other distribution $q^{(n)}$ under an alternative hypothesis H_2. The problem is to decide which hypothesis is true based on a sequence of random observations in a finite set $\mathcal{X}$. Let $A_n \subseteq \mathcal{X}^n$ be an acceptance region for the null hypothesis. Then, two probabilities of error can occur. The type-1 error probability is defined as

$$\alpha_n \triangleq p^{(n)}(A_n^c),$$

where A_n^c denotes the complement of A_n; α_n basically denotes the probability that H_2 is chosen given that H_1 is true. The type-2 error probability is defined as

$$\beta_n \triangleq q^{(n)}(A_n),$$

which denotes the probability of choosing H_1 when H_2 is true. In general, one wishes to minimize both probabilities, but there is a trade-off. Another approach, is to minimize one of the probabilities of error subject to a constraint on the other probability of error.

The best achievable error exponent for hypothesis testing has been thoroughly studied for independent and identically distributed (i.i.d.) sources and Markov sources, and the error exponents have been determined. The result for i.i.d. sources (known as Stein's Lemma) is given by the following theorem.

Proposition 3.1 (Stein's Lemma) [18], [21]: Let $\{X_1, X_2, \ldots\}$ be an i.i.d. source generated according to $p^{(n)}$ under H_1 and according to $q^{(n)}$ under H_2 with respective initial distributions p and q. Suppose that $D(p\|q) < \infty$. Let $A_n \subseteq \mathcal{X}^n$ be an acceptance region for H_1 and α_n and β_n denote the type-1 and type-2 error probabilities, respectively. For $\varepsilon \in (0, 1)$, define

$$\beta_n^\varepsilon \triangleq \min_{A_n \subseteq \mathcal{X}^n : \alpha_n < \varepsilon} \beta_n.$$

Then

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta_n^\varepsilon = D(p\|q).$$

The best achievable error exponent for testing between two irreducible Markov sources is given by the following theorem.

Proposition 3.2 [5]: Let $\{X_1, X_2, \ldots\}$ be a stationary and irreducible Markov source generated according to $p^{(n)}$ under H_1 and according to $q^{(n)}$ under H_2 with respective initial distributions p and q and respective probability transition matrices P and Q. Suppose that p and P are absolutely continuous with respect to q and Q respectively. Then

$$\lim_{n\to\infty} -\frac{1}{n}\log \beta_n^{\varepsilon} = \sum_{i\in\mathcal{X}} \pi_i \sum_{j\in\mathcal{X}} p_{ij}\log\frac{p_{ij}}{q_{ij}},$$

where $\pi = (\pi_1, \ldots, \pi_M)$ is the unique stationary distribution of P.

The proof involves large deviation theory. It mainly relies on Sanov's Theorem [12] for the *type* or *empirical transition-count matrix* of an arbitrary sample $x^n \in \mathcal{X}^n$ of the source. The type of x^n is the probability distribution on $\mathcal{X}^2$ giving mass $N(i,j,x^n)/n$ to $(i,j) \in \mathcal{X}^2$, where $N(i,j,x^n)$ denotes the number of transitions from i to j in x^n with the cyclic convention that x_1 follows x_n. The ij-th entry of the empirical transition-count matrix is also given by $N(i,j,x^n)/n$. Sanov's Theorem can be roughly described as follows. The probability of seeing sample sequences for which the type is far from the true distribution decreases to zero exponentially in the sample size. The decision region used in the proof is described as follows. Upon observing a sample from the source, choose $p^{(n)}$ as the true distribution iff the empirical transition-count matrix of the sample is "close" to the probability transition matrix P. Recently, in [15] the author generalizes Stein's Lemma for testing between arbitrary sources (not necessarily, Markov, stationary, ergodic, etc.) using an information spectrum approach. He obtained a lower bound and an upper bound to the error exponent which are not necessarily computable in general. In the sequel, we provide an alternative proof of the above proposition which follows along the same lines as in the proof of Proposition 3.1. Let us first show that the normalized log-likelihood ratio $\frac{1}{n}\log\frac{p^{(n)}(X^n)}{q^{(n)}(X^n)}$ converges to a limit with probability 1 under the null hypothesis.

Lemma 3.1 Let $\{X_1, X_2, \ldots, \}$ be a Markov source that is stationary and irreducible under both $p^{(n)}$ and $q^{(n)}$. Then

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{p^{(n)}(X^n)}{q^{(n)}(X^n)} = \sum_{i\in\mathcal{X}}\pi_i\sum_{j\in\mathcal{X}}p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

with probability 1 under $p^{(n)}$, where $\pi = (\pi_1, \ldots, \pi_M)$ is the unique stationary distribution of P.

Proof: Note that the normalized log-likelihood ratio can be written as

$$\frac{1}{n}\log\frac{p^{(n)}(X^n)}{q^{(n)}(X^n)} = \frac{1}{n}\log\frac{p(X_1)}{q(X_1)} + \frac{n-1}{n}\left[\frac{1}{n-1}\sum_{i=2}^{n}\log\frac{p(X_i|X_{i-1})}{q(X_i|X_{i-1})}\right].$$

In the limit, as $n \to \infty$, the first term approaches 0, and the second term which is the time average of $\log\frac{p(X_i|X_{i-1})}{q(X_i|X_{i-1})}$ approaches the statistical average with probability 1 under the probability distribution $p^{(n)}$, by the ergodic theorem [10, p. 13]. The statistical average of this quantity with respect to $p^{(n)}$ is

$$\begin{aligned} E\left(\log\frac{p(X_i|X_{i-1})}{q(X_i|X_{i-1})}\right) &= \sum_{x^n\in\mathcal{X}^n}p(x^n)\log\frac{p(x_i|x_{i-1})}{q(x_i|x_{i-1})} \\ &= \sum_{x_{i-1},x_i}p(x_{i-1},x_i)\log\frac{p(x_i|x_{i-1})}{q(x_i|x_{i-1})} \\ &= \sum_{i\in\mathcal{X}}\pi_i\sum_{j\in\mathcal{X}}p_{ij}\log\frac{p_{ij}}{q_{ij}}, \end{aligned}$$

where the last equality follows by stationarity; hence we obtain the desired result.

□

Remark: By the previous lemma and Theorem 3.1, the following holds with probability 1 under $p^{(n)}$.

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\frac{p^{(n)}(X^n)}{q^{(n)}(X^n)} &= \lim_{n\to\infty}\frac{1}{n}D(p^{(n)}\|q^{(n)}) \\ &= \sum_{i\in\mathcal{X}}\pi_i\sum_{j\in\mathcal{X}}p_{ij}\log\frac{p_{ij}}{q_{ij}},\end{aligned}$$

where $\pi = (\pi_1, \ldots, \pi_M)$ is the unique stationary distribution of P. We now provide a simple alternative proof for Proposition 3.2, which goes along the same lines as in [18, p. 309].

Proof of Proposition 3.2: We first construct a sequence of acceptance regions $A_n \in \mathcal{X}^n$ such that $\alpha_n < \varepsilon$ for n sufficiently large and

$$\lim_{n\to\infty} -\frac{1}{n}\log\beta_n = L,$$

where

$$L \triangleq \lim_{n\to\infty}\frac{1}{n}D(p^{(n)}\|q^{(n)}),$$

which exists by Theorem 3.1. Fix $\delta > 0$ and let

$$A_n = \left\{x^n \in \mathcal{X}^n : 2^{n(L-\delta)} \le \frac{p^{(n)}(x^n)}{q^{(n)}(x^n)} \le 2^{n(L+\delta)}\right\}.$$

Then $p^{(n)}(A_n) \to 1$ as $n \to \infty$. This follows from the previous remark. Hence, for $\delta = \varepsilon$ and sufficiently large n, $\alpha_n = p^{(n)}(A_n^c) < \varepsilon$. By definition of A_n, we have that

$$\begin{aligned}\beta_n = q^{(n)}(A_n) &= \sum_{x^n \in A_n} q^{(n)}(x^n) \\ &\leq \sum_{x^n \in A_n} p^{(n)}(x^n) 2^{-n(L-\delta)} \\ &= 2^{-n(L-\delta)} \sum_{x^n \in A_n} p^{(n)}(x^n) \\ &= 2^{-n(L-\delta)}(1-\alpha_n).\end{aligned}$$

Similarly, it can be shown that

$$\beta_n \geq 2^{-n(L+\delta)}(1-\alpha_n).$$

Hence,

$$-\frac{1}{n}\log\beta_n \geq L - \delta - \frac{1}{n}\log(1-\alpha_n),$$

and

$$-\frac{1}{n}\log\beta_n \leq L + \delta - \frac{1}{n}\log(1-\alpha_n).$$

Thus

$$\lim_{n\to\infty} -\frac{1}{n}\log\beta_n = L.$$

We now prove that no other sequence of acceptance regions does better. Let $B_n \subseteq \mathcal{X}^n$ be any other sequence of acceptance regions with type 1 error probability $\alpha_n' = p^{(n)}(B_n^c) < \varepsilon$, and type 2 error probability $\beta_n' = q^{(n)}(B_n)$. We will show that $\beta_n' \geq 2^{-n(L-\delta)}$, where $\delta > 0$ is arbitrary. We have the following.

$$\begin{aligned}
\beta'_n = q^{(n)}(B_n) &\geq q^{(n)}(A_n \cap B_n) \\
&= \sum_{x^n \in A_n \cap B_n} q^{(n)}(x^n) \\
&\geq \sum_{x^n \in A_n \cap B_n} p^{(n)}(x^n) 2^{-n(L+\delta)} \\
&= 2^{-n(L+\delta)} \sum_{x^n \in A_n \cap B_n} p^{(n)}(x^n) \\
&\geq (1 - \alpha_n - \alpha'_n) 2^{-n(L+\delta)},
\end{aligned}$$

where the last inequality follows from the union bound as follows:

$$\begin{aligned}
\sum_{x^n \in A_n \cap B_n} p^{(n)}(x^n) &= p^{(n)}(A_n \cap B_n) \\
&= 1 - p^{(n)}(A_n^c \cup B_n^c) \\
&\geq 1 - p^{(n)}(A_n^c) - p^{(n)}(B_n^c) \\
&= 1 - \alpha_n - \alpha'_n.
\end{aligned}$$

Hence

$$\frac{1}{n} \log \beta'_n \geq -L - \delta + \frac{1}{n} \log(1 - \alpha_n - \alpha'_n),$$

and since $\delta > 0$ is arbitrary,

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta'_n \leq L.$$

Thus, no sequence of sets B_n has an exponent larger than L. Since the sequence A_n achieves the exponent L, A_n is asymptotically optimal, and the best error exponent is L.

□

Remark: Our approach generalizes in a straightforward manner for stationary Markov sources that contain one irreducible essential class C_1 and an arbitrary number of inessential classes $C_2, \ldots, C_s$. Such a Markov source is said to be *indecomposable* [7]. In this case, the stationary distribution is $\pi = (\pi_1, 0, \ldots, 0)$, where π_1 is the stationary distribution corresponding to C_1 and the zeros correspond to inessential classes. We have the following result.

Corollary 3.4 Let $\{X_1, X_2, \ldots\}$ be a stationary Markov source generated according to $p^{(n)}$ under H_1 and according to $q^{(n)}$ under H_2 with respective probability transition matrices P and Q. Suppose that the Markov source has one essential class C_1 with j indices and an arbitrary number of inessential classes $C_2, \ldots, C_s$. Also, suppose that p and P are absolutely continuous with respect to q and Q respectively. Then

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta_n^{\varepsilon} = \sum_{i\in C_1} \pi_i \sum_{k\in C_1} p_{ik} \log \frac{p_{ik}}{q_{ik}},$$

where $\pi = (\pi_1, \ldots, \pi_j)$ is the unique stationary distribution corresponding to C_1.

Chapter 4

Rényi's Information Measure Rates for Finite-Alphabet Markov Sources

Let $\{X_1, X_2, \ldots\}$ be a first-order time-invariant Markov source with finite-alphabet $\mathcal{X} = \{1, \ldots, M\}$. Consider the following two different probability laws for this source. Under the first law,

$$Pr\{X_1 = i\} =: p_i\text{and} \quad Pr\{X_{k+1} = j | X_k = i\} =: p_{ij}, \quad i, j \in \mathcal{X},$$

so that

$$p^{(n)}(i^n) := Pr\{X_1 = i_1, \ldots, X_n = i_n\} = p_{i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}, \quad i_1, \ldots, i_n \in \mathcal{X},$$

while under the second law the initial probabilities are q_i, the transition probabilities are q_{ij}, and the n-tuple probabilities are $q^{(n)}$. Let $p = (p_1, \ldots, p_M)$ and

$q = (q_1, \ldots, q_M)$ denote the initial distributions under $p^{(n)}$ and $q^{(n)}$ respectively.

The Rényi divergence [52] of order α between two distributions $\hat{p}$ and $\hat{q}$ defined on $\mathcal{X}$ is given by

$$D_\alpha(\hat{p}\|\hat{q}) = \frac{1}{\alpha - 1} \log \left(\sum_{i \in \mathcal{X}} \hat{p}_i^\alpha \hat{q}_i^{1-\alpha} \right),$$

where $0 < \alpha < 1$. This definition can be extended to $\alpha > 1$ if all $\hat{q}_i > 0$. The base of the logarithm is arbitrary. Similarly, the Rényi entropy of order α for $\hat{p}$ is defined as

$$H_\alpha(\hat{p}) = \frac{1}{1 - \alpha} \log \left(\sum_{i \in \mathcal{X}} \hat{p}_i^\alpha \right),$$

where $\alpha > 0$ and $\alpha \neq 1$. As $\alpha \to 1$, the Rényi divergence approaches the Kullback-Leibler divergence (relative entropy) given by

$$D(\hat{p}\|\hat{q}) = \sum_{i \in \mathcal{X}} \hat{p}_i \log \frac{\hat{p}_i}{\hat{q}_i},$$

and the Rényi entropy approaches the Shannon entropy. The above generalized information measures and their subsequent variations [57] were originally introduced for the analysis of memoryless sources. One natural direction for further studies is the investigation of the Rényi divergence rate

$$\lim_{n \to \infty} \frac{1}{n} D_\alpha(p^{(n)}\|q^{(n)}),$$

where

$$D_\alpha(p^{(n)}\|q^{(n)}) = \frac{1}{\alpha - 1} \log \left(\sum_{i^n \in \mathcal{X}^n} [p^{(n)}(i^n)]^\alpha [q^{(n)}(i^n)]^{1-\alpha} \right),$$

and of the Rényi entropy rate

$$\lim_{n \to \infty} \frac{1}{n} H_\alpha(p^{(n)}),$$

where

$$H_\alpha(p^{(n)}) = \frac{1}{1-\alpha} \log \left(\sum_{i^n \in \mathcal{X}^n} [p^{(n)}(i^n)]^\alpha \right),$$

for sources with memory, in particular Markov sources. Nemetz addressed these problems in [44], where he evaluated the Rényi divergence rate $\lim_{n\to\infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)})$ between two Markov sources characterized by $p^{(n)}$ and $q^{(n)}$, respectively, under the restriction that the initial probabilities p and q are strictly positive (i.e., all p_i's and q_i's are strictly positive).

In this chapter, we provide a generalization of the Nemetz result by establishing a computable expression for the Rényi divergence rate between Markov sources with *arbitrary* initial distributions. We also investigate the questions of whether the Rényi divergence rate reduces to the Kullback-Leibler divergence rate as $\alpha \to 1$ and the interchangeability of limits between n and α as $n \to \infty$ and as $\alpha \to 0$. We provide sufficient (but not necessary) conditions on the underlying Markov source distributions $p^{(n)}$ and $q^{(n)}$ for which the interchangeability of limits as $n \to \infty$ and as $\alpha \to 1$ is valid. We also give an example of non-interchangeability of limits as $n \to \infty$ and as $\alpha \to 1$. We also show that the interchangeability of limits as $n \to \infty$ and $\alpha \to 0$ always holds.

We next address the computation and the existence of the Rényi entropy rate $\lim_{n\to\infty} \frac{1}{n} H_\alpha(p^{(n)})$ for a Markov source with distribution $p^{(n)}$ and examine its limits as $\alpha \to 0$ and as $\alpha \to 1$. We also establish an operational characterization for the Rényi entropy rate by extending the variable-length source coding theorem for memoryless sources in [13] to Markov sources.

4.1 Rényi Divergence Rate

4.1.1 First-order Markov Sources

We assume first that the Markov source $\{X_1, X_2, \ldots\}$ is of order one. Later, we generalize the results for an arbitrary order k. The joint distributions of the random variables $(X_1, \ldots, X_n)$ under $p^{(n)}$ and $q^{(n)}$ are given respectively by

$$p^{(n)}(i^n) := Pr\{X_1 = i_1, \ldots, X_n = i_n\} = p_{i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n},$$

and

$$q^{(n)}(i^n) := Pr\{X_1 = i_1, \ldots, X_n = i_n\} = q_{i_1} q_{i_1 i_2} \cdots q_{i_{n-1} i_n}.$$

Let

$$V(n, \alpha) = \sum_{i^n \in \mathcal{X}^n} [p^{(n)}(i^n)]^{\alpha} [q^{(n)}(i^n)]^{1-\alpha}.$$

Then

$$V(n, \alpha) = \sum p_{i_1}^{\alpha} q_{i_1}^{1-\alpha} p_{i_1 i_2}^{\alpha} q_{i_1 i_2}^{1-\alpha} \cdots p_{i_{n-1} i_n}^{\alpha} q_{i_{n-1} i_n}^{1-\alpha},$$

where the sum is over $i_1, \ldots, i_n \in \mathcal{X}$. Define a new matrix $R = (r_{ij})$ by

$$r_{ij} = p_{ij}^{\alpha} q_{ij}^{1-\alpha}, \quad i, j = 1, \ldots, M.$$

Also, define two new $1 \times M$ vectors $s = (s_1, \ldots, s_M)$ and $\mathbf{1}$ by

$$s_i = p_i^{\alpha} q_i^{1-\alpha}, \quad \mathbf{1} = (1, \ldots, 1).$$

Then clearly $D_\alpha(p^{(n)} \| q^{(n)})$ can be written as

$$D_\alpha(p^{(n)} \| q^{(n)}) = \frac{1}{\alpha - 1} \log s R^{n-1} \mathbf{1}^t, \tag{4.1}$$

where $\mathbf{1}^t$ denotes the transpose of the vector $\mathbf{1}$. Without loss of generality, we will herein assume that there exists at least one $i \in \{1, \ldots, M\}$ for which $s_i > 0$, because otherwise (i.e., if $s_i = 0\ \forall i$), $D_\alpha(p^{(n)} \| q^{(n)})$ is infinite. We also assume that $0 < \alpha < 1$; we can allow the case of $\alpha > 1$ if $q > 0$ and $Q > 0$. We obtain the following results.

Theorem 4.1 If the matrix R is irreducible, then the Rényi divergence rate between $p^{(n)}$ and $q^{(n)}$ is given by

$$\lim_{n\to\infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) = \frac{1}{\alpha - 1} \log \lambda,$$

where λ is the largest positive real eigenvalue of R, and $0 < \alpha < 1$. Furthermore, the same result holds for $\alpha > 1$ if $q > 0$ and $Q > 0$.

Proof: By Proposition 2.4, let λ be the largest positive real eigenvalue of R with associated positive right eigenvector $b > 0$. Then

$$R^{n-1} b = \lambda^{n-1} b. \tag{4.2}$$

Let $R^{n-1} = (r_{ij}^{(n-1)})$ and $b^t = (b_1, b_2, \ldots, b_M)$. Also, let $b_L = \min_{1 \le i \le M}(b_i)$ and $b_U = \max_{1 \le i \le M}(b_i)$. Thus $0 < b_L \le b_i \le b_U\ \forall i$. Let $R^{n-1}\mathbf{1}^t = y^t$ where $y = (y_1, \ldots, y_M)$. Then, by (4.2)

$$\lambda^{n-1} b_i = \sum_{j=1}^{M} r_{ij}^{(n-1)} b_j \le \sum_{j=1}^{M} r_{ij}^{(n-1)} b_U = b_U y_i, \quad \forall i = 1, \ldots, M.$$

Similarly, it can be shown that $\lambda^{n-1} b_i \ge b_L y_i$, $\forall i = 1, \ldots, M$. Therefore

$$\frac{b_i}{b_U} \le \frac{y_i}{\lambda^{n-1}} \le \frac{b_i}{b_L}, \quad \forall i = 1, \ldots, M. \tag{4.3}$$

Since $sR^{n-1}\mathbf{1}^t = \sum_{i=1}^{M} s_i y_i$, it follows directly from (4.3) that

$$\frac{\sum_i s_i b_i}{b_U} \le \frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}} \le \frac{\sum_i s_i b_i}{b_L},$$

or

$$\frac{1}{n}\log\left(\frac{\sum_i s_i b_i}{b_U}\right) \le \frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right) \le \frac{1}{n}\log\left(\frac{\sum_i s_i b_i}{b_L}\right). \tag{4.4}$$

Note that s_i, b_i, b_U, b_L do not depend on n. Therefore, by (4.4),

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right) = 0,$$

since it is upper and lower bounded by two quantities that approach 0 as $n \to \infty$. Hence

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\left(sR^{n-1}\mathbf{1}^t\right) &= \lim_{n\to\infty}\frac{1}{n}\log\lambda^{n-1} + \lim_{n\to\infty}\frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right)\\ &= \log\lambda,\end{aligned}$$

and thus

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) &= \lim_{n\to\infty}\frac{1}{n(\alpha-1)}\log\left(sR^{n-1}\mathbf{1}^t\right)\\ &= \frac{1}{\alpha-1}\log\lambda.\end{aligned}$$

□

Using the above theorem and the canonical form of R we prove the following general result.

Theorem 4.2 Let R_i, $i = 1, \ldots, g$, be the irreducible matrices along the diagonal of the canonical form of the matrix R as shown in Proposition 2.2. Write the vector s as

$$s = (\tilde{s}_1, \ldots, \tilde{s}_h, \tilde{s}_{h+1}, \ldots, \tilde{s}_g, s_{g+1}, \ldots, s_l),$$

where the vector $\tilde{s}_i$ corresponds to R_i, $i = 1, \ldots, g$. The scalars $s_{g+1}, \ldots, s_l$ correspond to non self-communicating classes.

- Let λ_k be the largest positive real eigenvalue of R_k for which the corresponding vector $\tilde{s}_k$ is different from the zero vector, $k = 1, \ldots, g$. Let $\lambda^\star$ be the maximum over these λ_k's. If $\tilde{s}_k = 0$, $\forall k = 1, \ldots, g$, then let $\lambda^\star = 0$.
- For each inessential class C_i with corresponding vector $\tilde{s}_i \neq 0$, $i = h+1, \ldots, g$ or corresponding scalar $s_i \neq 0$, $i = g+1, \ldots, l$, let λ_j be the largest positive real eigenvalue of R_j if class C_j is reachable from class C_i. Let $\lambda^\dagger$ be the maximum over these λ_j's. If $\tilde{s}_i = 0$ and $s_i = 0$ for every inessential class C_i, then let $\lambda^\dagger = 0$.

Let $\lambda = \max\{\lambda^\star, \lambda^\dagger\}$. Then the Rényi divergence rate is given by

$$\lim_{n\to\infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) = \frac{1}{\alpha - 1} \log \lambda,$$

where $0 < \alpha < 1$. Furthermore, the same result holds for $\alpha > 1$ if $q > 0$ and $Q > 0$.

Proof: By Proposition 2.4, let λ_i be the largest positive real eigenvalue of R_i with associated positive right eigenvector $\tilde{b}_i > 0$, $i = 1, \ldots, g$. Let

$$b^t = (\tilde{b}_1, \ldots, \tilde{b}_h, \tilde{b}_{h+1}, \ldots, \tilde{b}_g, 0, \ldots, 0),$$

where the zeros correspond to non self-communicating classes. By Proposition 2.2 we have that

$$R^{n-1} = \begin{bmatrix} R_1^{n-1} & \dots & 0 & 0 & \dots & 0 & \dots & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & R_h^{n-1} & 0 & \dots & 0 & \dots & \dots & 0 \\ R_{h+11}^{(n-1)} & \dots & R_{h+1h}^{(n-1)} & R_{h+1}^{n-1} & \dots & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ R_{g1}^{(n-1)} & \dots & R_{gh}^{(n-1)} & R_{gh+1}^{(n-1)} & \dots & R_g^{n-1} & \dots & \dots & 0 \\ R_{g+11}^{(n-1)} & \dots & R_{g+1h}^{(n-1)} & R_{g+1h+1}^{(n-1)} & \dots & R_{g+1g}^{(n-1)} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ R_{l1}^{(n-1)} & \dots & R_{lh}^{(n-1)} & R_{lh+1}^{(n-1)} & \dots & R_{lg}^{(n-1)} & R_{lg+1}^{(n-1)} & \dots & 0 \end{bmatrix}.$$

Then

$$\begin{aligned} sR^{n-1}b &= \sum_{i=1}^{g} \tilde{s}_i R_i^{n-1} \tilde{b}_i + \sum_{i=h+1}^{g} \tilde{s}_i \left(R_{i1}^{(n-1)} \tilde{b}_1 + \cdots + R_{ii-1}^{(n-1)} \tilde{b}_{i-1} \right) \\ &+ \sum_{i=g+1}^{l} s_i \left(R_{i1}^{(n-1)} \tilde{b}_1 + \cdots + R_{ig}^{(n-1)} \tilde{b}_g \right). \end{aligned}$$

Rewrite the vector $\mathbf{1}$ as

$$\mathbf{1} = (\tilde{1}_1, \dots, \tilde{1}_h, \tilde{1}_{h+1}, \dots, \tilde{1}_g, 1, \dots, 1),$$

where $\tilde{1}_i,\ i = 1, \dots, g$ correspond to essential and inessential self-communicating classes and the 1's correspond to non self-communicating classes.

Let $R^{n-1}\mathbf{1}^t = y^t$ where

$$y = (\tilde{y}_1, \dots, \tilde{y}_h, \tilde{z}_{h+1} + \tilde{y}_{h+1}, \dots, \tilde{z}_g + \tilde{y}_g, \tilde{z}_{g+1}, \dots, \tilde{z}_l),$$

and

$$\begin{aligned}
\tilde{y}_i &= R_i^{n-1}\tilde{1}_i^t, && i = 1, \ldots, g, \\
\tilde{z}_i &= \sum_{j=1}^{i-1} R_{ij}^{(n-1)}\tilde{1}_j^t, && i = h+1, \ldots, g, \qquad (4.5)\\
\tilde{z}_i &= \sum_{j=1}^{g} R_{ij}^{(n-1)}\tilde{1}_j^t + \sum_{j=g+1}^{i-1} R_{ij}^{(n-1)}, && i = g+1, \ldots, l.
\end{aligned}$$

Therefore

$$sR^{n-1}\mathbf{1}^t = \sum_{i=1}^{g} \tilde{s}_i\tilde{y}_i + \sum_{i=h+1}^{g} \tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l} s_i\tilde{z}_i. \qquad (4.6)$$

As in the proof of Theorem 4.1, since $R_i\tilde{b}_i = \lambda_i\tilde{b}_i$, we can write

$$R_i^{n-1}\tilde{b}_i = \lambda_i^{n-1}\tilde{b}_i \leq b_U\tilde{y}_i, \quad i = 1, \ldots, g,$$

where $b_U = \max_{1\leq i\leq g}(b_{U_i})$ and b_{U_i} is the largest component of $\tilde{b}_i$, $i = 1, \ldots, g$. Similarly,

$$R_i^{n-1}\tilde{b}_i = \lambda_i^{n-1}\tilde{b}_i \geq b_L\tilde{y}_i, \quad i = 1, \ldots, g,$$

where $b_L = \min_{1\leq i\leq g}(b_{L_i})$ and b_{L_i} is the smallest component of $\tilde{b}_i$, $i = 1, \ldots, g$. Therefore

$$\frac{\lambda_i^{n-1}\tilde{b}_i}{b_U} \leq \tilde{y}_i \leq \frac{\lambda_i^{n-1}\tilde{b}_i}{b_L}, \quad i = 1, \ldots, g.$$

Hence

$$\frac{1}{b_U}\sum_{i=1}^{g} \tilde{s}_i\lambda_i^{n-1}\tilde{b}_i \leq \sum_{i=1}^{g} \tilde{s}_i\tilde{y}_i \leq \frac{1}{b_L}\sum_{i=1}^{g} \tilde{s}_i\lambda_i^{n-1}\tilde{b}_i.$$

Therefore, by (4.6)

$$\begin{aligned}
\frac{1}{b_U}\sum_{i=1}^{g} \tilde{s}_i\lambda_i^{n-1}\tilde{b}_i + \sum_{i=h+1}^{g} \tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l} s_i\tilde{z}_i &\leq sR^{n-1}\mathbf{1}^t \\
&\leq \frac{1}{b_L}\sum_{i=1}^{g} \tilde{s}_i\lambda_i^{n-1}\tilde{b}_i + \sum_{i=h+1}^{g} \tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l} s_i\tilde{z}_i,
\end{aligned}$$

or

$$\frac{1}{n}\log\left(\frac{1}{b_U}\sum_{i=1}^{g}\tilde{s}_i\left(\frac{\lambda_i}{\lambda}\right)^{n-1}\tilde{b}_i+\frac{1}{\lambda^{n-1}}\left(\sum_{i=h+1}^{g}\tilde{s}_i\tilde{z}_i+\sum_{i=g+1}^{l}s_i\tilde{z}_i\right)\right)\leq\frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right) \tag{4.7}$$

and

$$\frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right)\leq\frac{1}{n}\log\left(\frac{1}{b_L}\sum_{i=1}^{g}\tilde{s}_i\left(\frac{\lambda_i}{\lambda}\right)^{n-1}\tilde{b}_i+\frac{1}{\lambda^{n-1}}\left(\sum_{i=h+1}^{g}\tilde{s}_i\tilde{z}_i+\sum_{i=g+1}^{l}s_i\tilde{z}_i\right)\right), \tag{4.8}$$

where λ is as defined in the statement of the theorem. Our goal is to show that $\frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right)$ converges to 0 as $n \to \infty$. Let us first examine its lower bound in (4.7). We will provide a simpler lower bound which converges to 0 as $n \to \infty$. We have the following three cases.

1. $\lambda = \lambda_i$ and $\tilde{s}_i \neq 0$ for some $i = 1, \ldots, g$. In this case

   $$\frac{1}{n}\log\left(\frac{1}{b_U}\sum_{i=1}^{g}\tilde{s}_i\left(\frac{\lambda_i}{\lambda}\right)^{n-1}\tilde{b}_i+\frac{1}{\lambda^{n-1}}\left(\sum_{i=h+1}^{g}\tilde{s}_i\tilde{z}_i+\sum_{i=g+1}^{l}s_i\tilde{z}_i\right)\right)\geq\frac{1}{n}\log\left(\frac{1}{b_U}\tilde{s}_i\tilde{b}_i\right)$$

   which clearly converges to 0 as $n \to \infty$.

2. $\lambda = \lambda_j$ for some $j = 1, \ldots, g$ and $\tilde{s}_i \neq 0$ for some $i = h+1, \ldots, g$ where the class C_j is reachable from class C_i. By equating the entries of R^{n-1} and $R^{n-2}R$, it follows directly that $R_{ij}^{(n-1)}$ is equal to $R_{ij}^{(n-2)}R_j$ plus a weighted sum of non-negative sub-matrices.[1] Hence $R_{ij}^{(n-1)} \geq R_{ij}^{(n-2)}R_j$. By induction on $n \geq 3$, it follows directly that $R_{ij}^{(n-1)} \geq R_{ij}R_j^{n-2}$. Therefore

   $$\frac{1}{n}\log\left(\frac{1}{b_U}\sum_{i=1}^{g}\tilde{s}_i\left(\frac{\lambda_i}{\lambda}\right)^{n-1}\tilde{b}_i+\frac{1}{\lambda^{n-1}}\left(\sum_{i=h+1}^{g}\tilde{s}_i\tilde{z}_i+\sum_{i=g+1}^{l}s_i\tilde{z}_i\right)\right)$$

[1] For example, if $R = \begin{bmatrix} R_1 & 0 \\ R_{21} & R_2 \end{bmatrix}$, then $R_{21}^{(n-1)} = R_{21}^{(n-2)}R_1 + R_2^{n-2}R_{21}$.

$$\begin{aligned}
&\geq \frac{1}{n}\log\left(\frac{1}{\lambda^{n-1}}\tilde{s}_i z_i\right) \\
&\geq \frac{1}{n}\log\left(\frac{1}{\lambda^{n-1}}\tilde{s}_i R_{ij}^{(n-1)}\tilde{1}_j^t\right) \qquad (4.9) \\
&\geq \frac{1}{n}\log\left(\frac{1}{\lambda^{n-1}}\tilde{s}_i R_{ij} R_j^{n-2}\tilde{1}_j^t\right) \qquad (4.10)
\end{aligned}$$

where (4.9) follows from (4.5). Using similar technique as in Theorem 4.1, it can be verified that the right-hand term of (4.10) converges to 0 as $n \to \infty$.

3. $\lambda = \lambda_j$ for some $j = 1, \ldots, g$ and $s_i \neq 0$ for some $i = g+1, \ldots, l$ where the class C_j is reachable from class C_i. The proof for this case is similar to that of case 2.

Let us now examine the upper bound to $\frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right)$ in (4.8). By definition of λ, it is obvious that $\frac{\lambda_i}{\lambda} \leq 1$, for all $i = 1, \ldots, g$ such that $\tilde{s}_i \neq 0$. Therefore

$$\frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right) \leq \frac{1}{n}\log\left(\frac{1}{b_L}\sum_{i=1}^{g}\tilde{s}_i\tilde{b}_i + \frac{1}{\lambda^{n-1}}\left(\sum_{i=h+1}^{g}\tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l}s_i\tilde{z}_i\right)\right). \qquad (4.11)$$

Note that

$$\begin{aligned}
\sum_{i=h+1}^{g}\tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l}s_i\tilde{z}_i \;=\; & \sum_{i=h+1}^{g}\sum_{j=1}^{i-1}\tilde{s}_i R_{ij}^{(n-1)}\tilde{1}_j^t + \sum_{i=g+1}^{l}\sum_{j=1}^{g}s_i R_{ij}^{(n-1)}\tilde{1}_j^t \\
& + \sum_{i=g+1}^{l}\sum_{j=g+1}^{i-1}s_i R_{ij}^{(n-1)}.
\end{aligned}$$

Our approach is to provide an upper bound to the bound in (4.11), simply by providing an upper bound on $R_{ij}^{(n-1)}$, $i = h+1, \ldots, l$, $j = 1, \ldots, g+1$. If $R_{ij}^{(n-1)} \neq 0$ for some n, then class C_j is reachable from class C_i (it is enough to check for $n = 2, \ldots, l$, since the number of classes is l). From the block form of R, if $R_{ij}^{(n-1)} \neq 0$, then it is a weighted sum involving products of powers of R_i and R_j (which are irreducible) and possibly some other sub-matrices (which are irreducible) along the diagonal[2] of R. By applying Proposition 2.6 to each of these irreducible sub-matrices if $\tilde{s}_i \neq 0$ or $s_i \neq 0$ (since $R_{ij}^{(n-1)}$ is multiplied by $\tilde{s}_i$ or s_i), $R_{ij}^{(n-1)}$ is upper bounded by linear combinations of powers of the largest eigenvalues of the sub-matrices along the diagonal of R for which $\tilde{s}_i \neq 0$, $i = h+1, \ldots, g$, or for which the corresponding class is reachable from class C_i, $i = g+1, \ldots, l$. For example, in the case of the R as given in the footnote, $R_{21}^{(n-1)} \leq \lambda^{n-2} D$, where $D > 0$ and its entries are independent of n. We have the following (here $g = l = 2$ and $h = 1$).

$$\begin{aligned}
&\frac{1}{n} \log \left(\frac{1}{b_L} \sum_{i=1}^{g} \tilde{s}_i \tilde{b}_i + \frac{1}{\lambda^{n-1}} \left(\sum_{i=h+1}^{g} \tilde{s}_i \tilde{z}_i + \sum_{i=g+1}^{l} s_i \tilde{z}_i \right) \right) \\
&= \frac{1}{n} \log \left(\frac{1}{b_L} \sum_{i=1}^{2} \tilde{s}_i \tilde{b}_i + \frac{1}{\lambda^{n-1}} \left(\tilde{s}_2 \tilde{z}_2 \right) \right) \\
&= \frac{1}{n} \log \left(\frac{1}{b_L} \sum_{i=1}^{2} \tilde{s}_i \tilde{b}_i + \frac{1}{\lambda^{n-1}} \left(\tilde{s}_2 R_{21}^{n-1} \tilde{1}_1^t \right) \right) \\
&\leq \frac{1}{n} \log \left(\frac{1}{b_L} \sum_{i=1}^{2} \tilde{s}_i \tilde{b}_i + \frac{1}{\lambda^{n-1}} \left(\lambda^{n-2} d \right) \right),
\end{aligned}$$

[2] For example, if $R = \begin{bmatrix} R_1 & 0 \\ R_{21} & R_2 \end{bmatrix}$, then $R_{21}^{(n-1)} = R_{21}^{(n-2)} R_1 + R_2^{n-2} R_{21}$. By induction, using the previous recursive formula, and Proposition 2.6, it is straightforward that $R_{21}^{(n-1)} \leq \lambda^{n-2} D$, where $D > 0$ and its entries are independent of n. Indeed, by Proposition 2.6, $R_2^{n-2} \leq \lambda_2^{n-2} D_2$, and $R_1 \leq \lambda_1 D_1$, where $D_2, D_1 > 0$ and their entries are independent of n. By induction, and by definition of λ, it follows that

$$R_{21}^{(n-1)} \leq \lambda^{n-2} D_3 + \lambda^{n-2} R_{21},$$

where $D_3 > 0$ and its entries are independent of n. Note also that R_{21} has entries independent of n. Hence, the desired result follows by taking $D = D_3 + R_{21}$.

where $d = \tilde{s}_2 D \tilde{1}_1^t$ is a positive constant. As $n \to \infty$, the above limit is obviously 0. Thus, the upper bound in (4.11) also converges to 0 as $n \to \infty$.

If R has three sub-matrices along the diagonal, then from the block form of R, the matrix $R_{31}^{(n-1)}$ is given recursively by the following formula. $R_{31}^{(n-1)} = R_{31}^{(n-2)} R_1 + R_{32}^{(n-2)} R_{21} + R_3^{n-2} R_{31}$. As in the previous example given in the footnote, by induction and Proposition 2.6, it is straightforward to show that $R_{31}^{(n-1)} \leq \lambda^{n-2} D_2 + \lambda^{n-3} D_3$, where $D_2, D_3 > 0$ and their entries are independent of n. In this case, by a reasoning similar to the previous example, it is straightforward to verify that

$$\sum_{i=h+1}^{g} \tilde{s}_i \tilde{z}_i + \sum_{i=g+1}^{l} s_i \tilde{z}_i$$

is upper bounded by

$$d_2 \lambda^{n-2} + d_3 \lambda^{n-3},$$

where d_2, d_3 are positive constants. Hence, the upper bound in (4.11) converges to 0 as $n \to \infty$. In general, using the fact that $R^n = R^{n-1} R$, a simple induction yields that

$$R_{ij}^{n-1} \leq d_2 \lambda^{n-2} + \cdots + d_l \lambda^{n-l},$$

for all $i = h+1, \ldots, l$, $j = 1, \ldots, g+1$, where l is the number of classes. Hence, the expression

$$\sum_{i=h+1}^{g} \tilde{s}_i \tilde{z}_i + \sum_{i=g+1}^{l} s_i \tilde{z}_i$$

is upper bounded by

$$d_2 \lambda^{n-2} + \cdots + d_l \lambda^{n-l},$$

where $d_2, \ldots, d_l$ are positive constants. Hence, from (4.11), we obtain the following.

$$\begin{aligned}
\frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right) &\leq \frac{1}{n}\log\left(\frac{1}{b_L}\sum_{i=1}^{g}\tilde{s}_i\tilde{b}_i + \frac{1}{\lambda^{n-1}}\left(\sum_{i=h+1}^{g}\tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l}s_i\tilde{z}_i\right)\right) \\
&\leq \frac{1}{n}\log\left(\frac{1}{b_L}\sum_{i=1}^{g}\tilde{s}_i\tilde{b}_i + \frac{1}{\lambda^{n-1}}\left(d_2\lambda^{n-2}+\cdots+d_l\lambda^{n-l}\right)\right) \\
&= \frac{1}{n}\log\left(\frac{1}{b_L}\sum_{i=1}^{g}\tilde{s}_i\tilde{b}_i + \frac{1}{\lambda}\left(d_2+\frac{d_3}{\lambda}+\cdots+\frac{d_l}{\lambda^{l-2}}\right)\right) \\
&= \frac{1}{n}\log d,
\end{aligned}$$

where d is a positive constant. Hence

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right) = 0, \tag{4.12}$$

since it is sandwiched between a lower bound (4.7) and an upper bound (4.8) that converge to 0 as $n \to \infty$. Finally, by (4.1) and (4.12), we get that

$$\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) = \frac{1}{\alpha-1}\log\lambda.$$

□

Observation 1: In [44], Nemetz showed that the Rényi divergence rate between two time-invariant Markov sources with *strictly positive* initial distributions is given by $\frac{1}{\alpha-1}\log\tilde{\lambda}$, where $\tilde{\lambda}$ is the largest positive real eigenvalue of R. The key tools used in establishing the Nemetz result [44] are Perron's formula and Perron-Frobenius theory for an arbitrary (not necessarily irreducible) non-negative matrix [32], [54]. The assumption that the initial distributions are strictly positive is essential, since as mentioned by Nemetz, the α-divergence rate is not necessarily continuous at points where the initial distributions vanish. In order to generalize the result for arbitrary initial distributions we used a different approach. We considered the canonical form of the matrix R and then used Perron-Frobenius theory on *each* irreducible sub-matrix along the diagonal of the canonical form instead of using Perron-Frobenius theory on the whole matrix at once. Although, the proof seems quite involved, the idea is very simple. As in Theorem 4.1, we employed a sandwich argument to show that the expression

$$\frac{1}{n}\log\left(\frac{sR^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right)$$

converges to 0 as $n \to \infty$ by showing that a lower bound and an upper bound converge to 0. The lower bound convergence is derived along the same lines as in Theorem 4.1. The key idea in deriving the convergence of the upper bound is to provide upper bounds to the sub-matrices off the diagonal of R^{n-1} which involve powers of positive eigenvalues of the irreducible sub-matrices along the diagonal of R. This is shown by induction with the aid of Proposition 2.6 applied to each of the irreducible sub-matrices along the diagonal of R^{n-1}. It is clear from our proof that no assumption of positivity is required on the initial distributions.

Observation 2: Note that by Theorem 4.2, the Rényi divergence rate between Markov sources with *arbitrary* initial distributions is not necessarily equal to $\frac{1}{\alpha-1}\log\tilde{\lambda}$, where $\tilde{\lambda}$ is the largest positive real eigenvalue of R. However, if the initial distributions are strictly positive, which implies directly that $\mathbf{s} > 0$, then Theorem 4.2 reduces to the Nemetz result. This follows directly from the fact that, in this case, $\lambda = \lambda^\star = \max\{\lambda_k\}$, $k = 1, \ldots, g$, and the fact that the determinant of a block lower triangular matrix is equal to the product of the determinants of the sub-matrices along the diagonal (thus the largest eigenvalue of this matrix is given by $\max\{\lambda_k\}$).

Theorem 4.3 The rate of convergence of the α-divergence rate between $p^{(n)}$ and $q^{(n)}$ is of the order $1/n$.

Proof: Note first that if $p^{(n)}$ and $q^{(n)}$ are irreducible, then by (4.4), the rate of convergence of the α-divergence rate is clearly of the order $1/n$ since s_i, b_i, B_U, b_L do not depend on n. For arbitrary $p^{(n)}$ and $q^{(n)}$ (not necessarily irreducible, stationary, etc.), from the proof of Theorem 4.2, it follows directly that the rate of convergence is also of the order $1/n$.

□

4.1.2 k-th Order Markov Sources

Now, suppose that the Markov source has an arbitrary order k. Define $\{W_n\}$ as the process obtained by k-step blocking the Markov source $\{X_n\}$; i.e.,

$$W_n \triangleq (X_n, X_{n+1}, \ldots, X_{n+k-1}).$$

Then

$$Pr(W_n = w_n | W_{n-1} = w_{n-1}, \ldots, W_1 = w_1) = Pr(W_n = w_n | W_{n-1} = w_{n-1}),$$

and $\{W_n\}$ is a first order Markov source with M^k states. Let $p_{w_{n-1}w_n} \triangleq Pr(W_n = w_n | W_{n-1} = w_{n-1})$. We next write the joint distributions of $\{X_n\}$ in terms of the conditional probabilities of $\{W_n\}$. For $n \geq k$, $V(n, \alpha)$, as defined before, is given by

$$V(n, \alpha) = \sum p_{w_1}^{\alpha} q_{w_1}^{1-\alpha} p_{w_1 w_2}^{\alpha} q_{w_1 w_2}^{1-\alpha} \cdots p_{w_{n-k} w_{n-k+1}}^{\alpha} q_{w_{n-k} w_{n-k+1}}^{1-\alpha},$$

where the sum is over $w_1, w_2, \ldots, w_{n-k+1} \in \mathcal{X}^k$. For simplicity of notation, let $(p_1, \ldots, p_{M^k})$ and $(q_1, \ldots, q_{M^k})$ denote the arbitrary initial distributions of W_1 under $p^{(n)}$ and $q^{(n)}$ respectively. Also let p_{ij} and q_{ij} denote the transition probability that W_n goes from index i to index j under $p^{(n)}$ and $q^{(n)}$ respectively, $i, j = 1, \ldots, M^k$. Define a new matrix $R = (r_{ij})$ by

$$r_{ij} = p_{ij}^{\alpha} q_{ij}^{1-\alpha}, \quad i, j = 1, \ldots, M^k. \tag{4.13}$$

Also, define two new $1 \times M^k$ vectors $\mathbf{s} = (s_1, \ldots, s_{M^k})$ and $\mathbf{1}$ by

$$s_i = p_i^{\alpha} q_i^{1-\alpha}, \quad \mathbf{1} = (1, \ldots, 1).$$

Then clearly $D_\alpha(p^{(n)} \| q^{(n)})$ can be written as

$$D_\alpha(p^{(n)} \| q^{(n)}) = \frac{1}{\alpha - 1} \log \mathbf{s} R^{n-k} \mathbf{1}^t,$$

where $\mathbf{1}^t$ denotes the transpose of the vector $\mathbf{1}$. It follows directly that with the new matrix R as defined in (4.13), all the previous results also hold for a Markov source of arbitrary order.

4.1.3 Numerical Examples

In this section, we use the natural logarithm.

Example 1: Let P and Q be two possible probability transition matrices for a first order Markov source $\{X_1, X_2, \ldots\}$ defined as follows:

$$P = \begin{bmatrix} 1/4 & 3/4 & 0 & 0 & 0 \\ 1/3 & 2/3 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/5 & 4/5 & 0 \\ 0 & 1/6 & 1/2 & 0 & 1/3 \end{bmatrix}, \quad Q = \begin{bmatrix} 1/5 & 4/5 & 0 & 0 & 0 \\ 1/6 & 5/6 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 3/4 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/3 & 0 & 1/6 \end{bmatrix}.$$

Note that P and Q are not irreducible. Indeed, P and Q have two essential classes and 1 inessential self-communicating class. Let the parameter $\alpha = 1/3$. The largest eigenvalues of the three sub-matrices along the diagonal of R are respectively: $\lambda_1 = 0.98676$, $\lambda_2 = 0.95937$, and $\lambda_3 = 0.20998$. Let $p = (0, 0, 3/4, 1/4, 0)$ and $q = (0, 0, 1/3, 2/3, 0)$ be two possible initial distributions under $p^{(n)}$ and $q^{(n)}$ respectively. It is straightforward to check that $p^{(n)}$ and $q^{(n)}$ are not stationary. For these given initial distributions, we get by Theorem 4.2 that $\lambda^\star = \lambda_2$ and $\lambda^\dagger = 0$. Therefore, the Rényi divergence rate is $\ln(\lambda_2)/(\alpha - 1) = 0.0622$. Note that λ_2 is *not* the largest eigenvalue of R. We also obtain the following.

n	$\frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)})$
10	0.0686
50	0.0635
100	0.0628
1000	0.06227
2000	0.06224
3000	0.06223

Clearly, as n gets large $\frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)})$ is closer to the Rényi divergence rate. Note however that, in general, the function $\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$ is not monotonic in n. Suppose that $\mathbf{s}$ has zero components on the first two classes. For example, let $p = (0, 1/4, 1/4, 0, 1/2)$ and $q = (1/4, 0, 0, 1/4, 1/2)$. In this case, $\lambda^\star = \lambda_3$, and $\lambda^\dagger = \max\{\lambda_1, \lambda_2\}$ (the first and second classes are reachable from the third). Therefore, the Rényi divergence rate is $\ln(\lambda_1)/(\alpha - 1) = 0.0199$. We also get the following.

n	$\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$
10	0.1473
50	0.0570
100	0.0413
1000	0.02223
2000	0.02111
3000	0.02074

Clearly, as n gets large $\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$ is closer to the Rényi divergence rate.

Suppose now that $\mathbf{s}$ has strictly positive components (as required in the Nemetz result). For example, let $p = (1/8, 1/4, 1/8, 1/4, 1/4)$ and $q = (1/10, 3/10, 2/10, 2/10, 2/10)$. In this case, $\lambda^\star = \lambda^\dagger = \max\{\lambda_1, \lambda_2, \lambda_3\} = \lambda_1$. Therefore, the Rényi divergence rate is $\ln(\lambda_1)/(\alpha - 1) = 0.01999$. Note that λ_1 is the largest eigenvalue of R which is expected since the components of $\mathbf{s}$ are strictly positive. We also get the following.

n	$\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$
10	0.0384
50	0.0343
100	0.0297
1000	0.02105
2000	0.02052
3000	0.02034

Clearly, as n gets large $\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$ is closer to the Rényi divergence rate.

Example 2: Suppose that the Markov source is of order 2 under $p^{(n)}$ and $q^{(n)}$ respectively. Let $\{W_1, W_2, \ldots\}$ be the process obtained by 2-step blocking the Markov source. Let P and Q be two possible transition matrices for $\{W_1, W_2, \ldots\}$ defined as follows:

$$P = \begin{bmatrix} 1/4 & 3/4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 3/5 & 2/5 & 0 & 0 \\ 0 & 0 & 1/5 & 4/5 \end{bmatrix},$$

and

$$Q = \begin{bmatrix} 2/3 & 1/3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 7/8 & 1/8 & 0 & 0 \\ 0 & 0 & 5/6 & 1/6 \end{bmatrix}.$$

Note that both P and Q are not irreducible. The set of indices $\{1, 2, 3\}$ forms an essential class, while the singleton set $\{4\}$ forms an inessential self-communicating class. Let the parameter $\alpha = 0.5$. The largest positive real eigenvalues of the two sub-matrices along the diagonal of R are respectively: $\lambda_1 = 0.9467$, $\lambda_2 = 0.3651$. Let $p = (1/4, 3/4, 0, 0)$ and $q = (1/5, 4/5, 0, 0)$ denote two possible initial distributions of W_1 under $p^{(n)}$ and $q^{(n)}$ respectively. Note that $p^{(n)}$ and $q^{(n)}$ are not stationary. For these given initial distributions, we get by Theorem 4.2 that $\lambda^* = \lambda_1$ and $\lambda^\dagger = 0$. Therefore, the Rényi divergence rate is $\ln(\lambda_1)/(\alpha - 1) = 0.1095$. We also obtain the following.

n	$\frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)})$
10	0.0817
50	0.1039
100	0.1066

Clearly, as n gets large $\frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)})$ is closer to the Rényi divergence rate.

Let us now suppose that $p = (1/4, 0, 0, 3/4)$ and $q = (1/3, 0, 0, 2/3)$. For these given initial distributions, we get by Theorem 4.2 that $\lambda^* = \lambda_1$ and $\lambda^\dagger = \lambda_1$. Therefore, the Rényi divergence rate is $\ln(\lambda_1)/(\alpha - 1) = 0.1095$. We also obtain the following.

n	$\frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)})$
10	0.1389
50	0.1153
100	0.1123

Clearly, as n gets large $\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$ is closer to the Rényi divergence rate.

4.2 Interchangeability of Limits

4.2.1 Limit as $\alpha \to 1$

We herein show that although the Rényi divergence reduces to the Kullback-Leibler divergence as $\alpha \to 1$, the Rényi divergence rate does not necessarily reduce to the Kullback-Leibler divergence rate. Without loss of generality, we will herein deal with first-order Markov sources since any k-th order Markov source can be converted to a first-order Markov source by k-step blocking it. We first show the following lemma.

Lemma 4.1 Let $A = (a_{ij})$ be an $n \times n$ matrix of rank $n-1$ with the property that $\sum_j a_{ij} = 0$ for each i. Define c_i to be the cofactor of a_{ii}; i.e., the determinant of the matrix obtained from A by deleting the i-th row and the i-th column and let $c = (c_1, c_2, \ldots, c_n)$. Then c is a non-zero vector and satisfies $cA = 0$.

Proof:

Step 1: First we prove that $c \neq 0$. The first $n-1$ columns of A are linearly independent, because otherwise, the rank of A is less than or equal to $n-2$ since the sum of the columns of A is 0. Thus there is at least one non-zero determinant Δ of size $(n-1)\times(n-1)$ which can be formed by deleting one row and the n-th column of A which follows from the fact that the determinant of a matrix is 0 iff the columns are linearly dependent. Let the deleted row be the k-th row. If $k = n$, $\Delta = c_n$ and so

$c \neq 0$. If $k < n$, add all the columns except the n-th column to the k-th column; this does not change the value of the determinant Δ. Because $\sum_j a_{ij} = 0$, the elements of the k-th column are now $-a_{1n}, -a_{2n}, \ldots, -a_{nn}$. Multiply the elements of this column by -1 and move this column to the rightmost position. This yields a new determinant with value $\pm\Delta$ because these operations affect only the sign of the determinant. However, the new determinant is just c_k, so that once again, $c \neq 0$. Thus at least one of the cofactors c_i is non-zero. Without loss of generality assume that $c_n \neq 0$. Next we prove that $cA = 0$.

Step 2: Consider the $n - 1$ equations

$$\sum_{i=1}^{n} a_{ij} x_i = 0 \qquad j \in \{1, 2, \ldots, n-1\}. \tag{4.14}$$

Note that $\sum_{i=1}^{n} a_{ij} x_i = 0$ is equivalent to $\sum_{i=1}^{n-1} a_{ij} x_i = -a_{nj} x_n$. Since $c_n \neq 0$, we can use Cramer's rule [41, p. 60] to solve these equations for $x_1, \ldots, x_{n-1}$ in terms of x_n as follows:

$$x_k = -x_n \frac{D_k}{c_n}, \tag{4.15}$$

where

$$D_k = \begin{vmatrix} a_{11} & a_{21} & \cdots & a_{k-1,1} & a_{n1} & a_{k+1,1} & \cdots & a_{n-1,1} \\ a_{12} & a_{22} & \cdots & a_{k-1,2} & a_{n2} & a_{k+1,2} & \cdots & a_{n-1,2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{1,n-1} & a_{2,n-1} & \cdots & a_{k-1,n-1} & a_{n,n-1} & a_{k+1,n-1} & \cdots & a_{n-1,n-1} \end{vmatrix},$$

and the elements from the n-th column have replaced the elements of the k-th column. If we add the other rows to the k-th row (note that the determinants are transposed here) and use the fact that $\sum_j a_{ij} = 0$ we get a new k-th row

$$-a_{1n}, -a_{2n}, \ldots, -a_{k-1,n}, -a_{nn}, -a_{k+1,n}, \ldots, -a_{n-1,n}.$$

After moving the k-th row and the k-th column to the last row and column position respectively, it follows that $D_k = -c_k$. From (4.15), if we put $x_n = c_n$, then $x_k = c_k$ for all $k \in \{1, 2, \ldots, n\}$. Because $\sum_j a_{ij} = 0$, any solution of (4.14) is a solution of the same equation for $j = n$. Thus $c = (c_1, \ldots, c_n)$ satisfies $cA = 0$.

□

Remark: A direct consequence of the above lemma generalizes Proposition 2.16 from ergodic Markov sources to irreducible Markov sources; this is achieved by setting $A = P - I$, where P is stochastic irreducible, and I is the identity matrix with the same dimension.

We next prove the following theorem.

Theorem 4.4 Given that $\alpha \in (0, 1)$, consider a Markov source $\{X_1, X_2, \ldots\}$ with two possible distributions $p^{(n)}$ and $q^{(n)}$ on $\mathcal{X}^n$. Let P and Q be the probability transition matrices associated with $p^{(n)}$ and $q^{(n)}$ respectively. Suppose that P and Q are irreducible and that P is absolutely continuous with respect to Q. Also, suppose that p is absolutely continuous with respect to q. Then

$$\begin{aligned}\lim_{\alpha \uparrow 1} \lim_{n \to \infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) &= \lim_{n \to \infty} \lim_{\alpha \uparrow 1} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) \\ &= \sum_{i,j} \pi_i p_{ij} \log \frac{p_{ij}}{q_{ij}},\end{aligned}$$

and therefore, the Rényi divergence rate reduces to the Kullback-Leibler divergence rate as $\alpha \uparrow 1$.

Proof: Under the above assumptions, the matrix R (as defined in Subsection 4.1.1) is irreducible. For convenience of notation denote the largest positive real eigenvalue of R by $\lambda(\alpha, R)$. We know by Proposition 2.8 that each eigenvalue of R is a continuous function of elements of R. Note that $R \to P$ as $\alpha \uparrow 1$, and the largest positive real eigenvalue of the stochastic matrix P is 1. Hence

$$\lim_{\alpha \uparrow 1} \lambda(\alpha, R) = 1.$$

Let a denote an arbitrary base of the logarithm. Then, by l'Hôpital's rule, we find that

$$\lim_{\alpha \uparrow 1} \frac{\log \lambda(\alpha, R)}{\alpha - 1} = \frac{1}{\ln a} \lambda'(1, R) \triangleq \frac{1}{\ln a} \left. \frac{\partial \lambda(\alpha, R)}{\partial \alpha} \right|_{\alpha=1} \tag{4.16}$$

which is well defined by Proposition 2.9 since the algebraic multiplicity of $\lambda(\alpha, R)$ is 1 (R is irreducible) by Proposition 2.7. The equation defining the largest positive eigenvalue $\lambda(\alpha, R) = \lambda$ of R is

$$\begin{vmatrix} p_{11}^{\alpha} q_{11}^{1-\alpha} - \lambda & p_{12}^{\alpha} q_{12}^{1-\alpha} & \cdots & p_{1M}^{\alpha} q_{1M}^{1-\alpha} \\ p_{21}^{\alpha} q_{21}^{1-\alpha} & p_{22}^{\alpha} q_{22}^{1-\alpha} - \lambda & \cdots & p_{2M}^{\alpha} q_{2M}^{1-\alpha} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M1}^{\alpha} q_{M1}^{1-\alpha} & p_{M2}^{\alpha} q_{M2}^{1-\alpha} & \cdots & p_{MM}^{\alpha} q_{MM}^{1-\alpha} - \lambda \end{vmatrix} = 0, \tag{4.17}$$

where $M = |\mathcal{X}|$. By Lemma 2.4, differentiating this equation with respect to α, we get that

$$D_1 + D_2 + \cdots + D_M = 0, \tag{4.18}$$

where D_i is the determinant obtained from (4.17) by replacing the i-th row by

$$(p_{i1}^{\alpha} q_{i1}^{1-\alpha} \ln \frac{p_{i1}}{q_{i1}}, \dots, p_{ii}^{\alpha} q_{ii}^{1-\alpha} \ln \frac{p_{ii}}{q_{ii}} - \lambda'(\alpha), \dots, p_{iM}^{\alpha} q_{iM}^{1-\alpha} \ln \frac{p_{iM}}{q_{iM}}).$$

and leaving the other $M-1$ rows unchanged. In this equation, λ' denotes the derivative of λ with respect to α. Note that if we add in D_i all the other columns to the i-th column, the value of the determinant remains unchanged. Therefore, for $\alpha = 1$ and hence $\lambda = 1$, D_i is the determinant

$$\begin{vmatrix} p_{11} - 1 & \dots & 0 & \dots & p_{1M} \\ p_{21} & \dots & 0 & \dots & p_{2M} \\ \vdots & \vdots & 0 & \dots & \vdots \\ p_{i-1,1} & \dots & 0 & \dots & p_{i-1,M} \\ p_{i1} \ln \frac{p_{i1}}{q_{i1}} & \dots & S(X|i) - \lambda' & \dots & p_{iM} \ln \frac{p_{iM}}{q_{iM}} \\ p_{i+1,1} & \dots & 0 & \dots & p_{i+1,M} \\ \vdots & \vdots & 0 & \dots & \vdots \\ p_{M1} & \dots & 0 & \dots & p_{MM} - 1 \end{vmatrix},$$

where

$$S(X|i) = \sum_{j=1}^{M} p_{ij} \ln \frac{p_{ij}}{q_{ij}}.$$

A zero occurs in all the entries of the i-th column except for the i-th entry, since $\sum_{j=1}^{M} p_{lj} = 1$. We conclude that

$$D_i = (S(X|i) - \lambda'(1))c_i, \tag{4.19}$$

where c_i is the $M - 1 \times M - 1$ cofactor of $p_{ii} - 1$ in the above determinant for the case $\alpha = 1$, given by

$$c_i = \begin{vmatrix} p_{11}-1 & \dots & p_{1,i-1} & \dots & p_{1M} \\ p_{21} & \dots & p_{2,i-1} & \dots & p_{2M} \\ \vdots & \dots & \dots & \dots & \vdots \\ p_{i-1,1} & \dots & p_{i-1,i-1}-1 & \dots & p_{i-1,M} \\ p_{i+1,1} & \dots & p_{i+1,i-1} & \dots & p_{i+1,M} \\ \vdots & \dots & \dots & \dots & \vdots \\ p_{M1} & \dots & p_{M,i-1} & \dots & p_{MM}-1 \end{vmatrix}.$$

After substituting (4.19) in (4.18) and solving for $\lambda'(1)$, we obtain by (4.16) that

$$\lim_{\alpha \uparrow 1} \frac{\log \lambda(\alpha, R)}{\alpha - 1} = \frac{1}{\ln a}\lambda'(1, R) = \frac{1}{\ln a}\sum_{i=1}^{M} \pi_i S(X|i), \qquad (4.20)$$

where

$$\pi_i = \frac{c_i}{\sum_j c_j}.$$

As $\alpha \uparrow 1$, $R \to P$; let $A = P - I$. Since the stationary distribution of the irreducible matrix R is unique, the rank of A is $n - 1$ because the nullity of A is 1 in this case. Hence, the conditions in Lemma 4.1 are satisfied. Therefore, $cA = 0$, which is equivalent to $cP = c$. Note that c is the non-normalized stationary distribution of P and (4.20) is just the Kullback-Leibler divergence rate between P and Q by Theorem 3.1. □

For the case $\alpha \in (1, \infty)$, we can obtain a similar result under the conditions that the matrix Q and the initial distribution q are positive. This is stated in the following corollary (whose proof is identical to the proof of Theorem 4.4).

Corollary 4.1 Given that $\alpha \in (1, \infty)$, consider a Markov source $\{X_1, X_2, \ldots\}$ with two possible distributions $p^{(n)}$ and $q^{(n)}$ on $\mathcal{X}^n$. Let P and Q be the probability transition matrices associated with $p^{(n)}$ and $q^{(n)}$ respectively. If the matrix P is irreducible, the matrix Q is positive, and the initial distribution q with respect to $q^{(n)}$ is positive, then

$$\begin{aligned}\lim_{\alpha\downarrow 1}\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) &= \lim_{n\to\infty}\lim_{\alpha\downarrow 1}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) \\ &= \sum_{i,j}\pi_i p_{ij}\log\frac{p_{ij}}{q_{ij}},\end{aligned}$$

and therefore, the Rényi divergence rate reduces to the Kullback-Leibler divergence rate as $\alpha \downarrow 1$.

The following example illustrates that the Rényi divergence rate does not necessarily reduce to the Kullback-Leibler divergence rate if the conditions of the previous theorem are not satisfied.

Example: Given that $\alpha \in (0, 1) \cup (1, \infty)$, let P and Q be the following:

$$P = \begin{bmatrix} 1/4 & 3/4 & 0 \\ 3/4 & 1/4 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad Q = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

Suppose that $p^{(n)}$ is stationary with stationary distribution $(b/2, b/2, 1-b)$, where $0 < b < 1$ is arbitrary. Also, suppose that the initial distribution q is positive. By Theorem 3.2, a simple computation yields that the Kullback-Leibler divergence rate is given by $\log_2 3 - 2b + (3b/4)\log_2 3$, where the logarithm is to the base 2. The eigenvalues of R are: $\lambda_1 = 1/(3^{1-\alpha})$, $\lambda_2 = 4^{-\alpha}/(3^{1-\alpha}) + 4^{-\alpha}/(3^{1-2\alpha})$, and $\lambda_3 =$

$4^{-\alpha}/(3^{1-\alpha}) - 4^{-\alpha}/(3^{1-2\alpha})$. Note that $\mathbf{s} > 0$ and that, if $0 < \alpha < 1$, $\max_{1 \le i \le 3}\{\lambda_i\} = \lambda_2$. By Theorem 4.2, the Rényi divergence rate is $(\alpha - 1)^{-1} \log_2 \lambda_2$. By l'Hôpital's rule, we get that $\lim_{\alpha \uparrow 1}(\alpha - 1)^{-1} \log_2 \lambda_2 = (7/4) \log_2 3 - 2$. Therefore

$$\lim_{\alpha \uparrow 1} \lim_{n \to \infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) = (7/4) \log_2 3 - 2.$$

On the other hand, if $\alpha > 1$, $\max_{1 \le i \le 3}\{\lambda_i\} = \lambda_1$. Therefore, the Rényi divergence rate is given by $(\alpha - 1)^{-1} \log_2 \lambda_1$. Clearly, $\lim_{\alpha \downarrow 1}(\alpha - 1)^{-1} \log_2 \lambda_1 = \log_2 3$. Hence

$$\lim_{\alpha \downarrow 1} \lim_{n \to \infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) = \log_2 3.$$

Therefore, the interchangeability of limits is not valid since

$$\lim_{\alpha \uparrow 1} \lim_{n \to \infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) < \lim_{n \to \infty} \lim_{\alpha \to 1} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) < \lim_{\alpha \downarrow 1} \lim_{n \to \infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}).$$

4.2.2 Limit as $\alpha \downarrow 0$

We obtain the following result.

Theorem 4.5 Let $\alpha \in (0, 1)$. Consider a Markov source $\{X_1, X_2, \ldots\}$ with two possible distributions $p^{(n)}$ and $q^{(n)}$ on $\mathcal{X}^n$. Let P and Q be the probability transition matrices on $\mathcal{X}$ associated with $p^{(n)}$ and $q^{(n)}$, respectively. Then

$$\lim_{\alpha \downarrow 0} \lim_{n \to \infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) = \lim_{n \to \infty} \lim_{\alpha \downarrow 0} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}).$$

Proof: By Theorem 4.2, we have

$$\lim_{n \to \infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) = \frac{1}{\alpha - 1} \log \lambda(\alpha, R).$$

By Proposition 2.8, $\lambda(\alpha, R) \to \lambda(0, R)$ as $\alpha \downarrow 0$. Hence

$$\lim_{\alpha \downarrow 0} \lim_{n \to \infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) = -\log \lambda(0, R).$$

On the other hand

$$\lim_{\alpha \downarrow 0} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) = \frac{1}{n} \log \hat{\mathbf{s}} Y \mathbf{1}^t,$$

where $\hat{\mathbf{s}} = \lim_{\alpha \downarrow 0} \mathbf{s}$ and $Y = \lim_{\alpha \downarrow 0} R$. Therefore by again applying Theorem 4.2 to Y we get

$$\lim_{n \to \infty} \lim_{\alpha \downarrow 0} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) = -\log \lambda(0, R).$$

Hence the interchangeability of limits is always valid between n and α as $n \to \infty$ and as $\alpha \downarrow 0$. □

4.3 Rényi's Entropy Rate

The existence and the computation of the Rényi entropy rate of a Markov source can be deduced from the existence and the computation of the Rényi divergence rate. Indeed, if $q^{(n)}$ is stationary memoryless with uniform marginal distribution then for any $\alpha > 0$, $\alpha \neq 1$,

$$D_\alpha(p^{(n)} \| q^{(n)}) = n \log M - H_\alpha(p^{(n)}).$$

Therefore

$$\lim_{n \to \infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) = \log M - \lim_{n \to \infty} \frac{1}{n} H_\alpha(p^{(n)}). \tag{4.21}$$

Hence, the existence and the computation of the Rényi entropy rate follows directly from Theorem 4.1 if the Markov source is irreducible, and from Theorem 4.2 if the

Markov source is arbitrary (not necessarily irreducible). Actually, $\lim_{n\to\infty}\frac{1}{n}H_\alpha(p^{(n)})$ can be computed directly from Theorem 4.1 or from Theorem 4.2 by determining λ with $R=(p_{ij}^\alpha)$ and $s_i=p_i^\alpha$, and setting $\lim_{n\to\infty}\frac{1}{n}H_\alpha(p^{(n)})=\frac{1}{1-\alpha}\log\lambda$. A formula for the Rényi entropy rate was established earlier in [46] and [47], but only for the particular case of ergodic Markov sources. We have the following corollaries. The proof follows along the same lines as for the Rényi divergence rate or by using (4.21) with $q^{(n)}$ stationary memoryless and uniformly distributed.

Corollary 4.2 If the Markov source under $p^{(n)}$ is irreducible, then the Rényi entropy rate is given by

$$\lim_{n\to\infty}\frac{1}{n}H_\alpha(p^{(n)})=\frac{1}{1-\alpha}\log\lambda,$$

where λ is the largest positive real eigenvalue of R, and $0<\alpha$, $\alpha\neq 1$.

Corollary 4.3 Let R_i, $i=1,\ldots,g$, be the irreducible matrices along the diagonal of the canonical form of the matrix R as shown in Proposition 2.2. Write the vector s as

$$s=(\tilde{s}_1,\ldots,\tilde{s}_h,\tilde{s}_{h+1},\ldots,\tilde{s}_g,s_{g+1},\ldots,s_l),$$

where the vector $\tilde{s}_i$ corresponds to R_i, $i=1,\ldots,g$. The scalars $s_{g+1},\ldots,s_l$ correspond to non self-communicating classes.

- Let λ_k be the largest positive real eigenvalue of R_k for which the corresponding vector $\tilde{s}_k$ is different from the zero vector, $k=1,\ldots,g$. Let $\lambda^\star$ be the maximum over these λ_k's. If $\tilde{s}_k=0$, $\forall k=1,\ldots,g$, then let $\lambda^\star=0$.

- For each inessential class C_i with corresponding vector $\tilde{s}_i \neq 0$, $i = h+1, \ldots, g$ or corresponding scalar $s_i \neq 0$, $i = g+1, \ldots, l$, let λ_j be the largest positive real eigenvalue of R_j if class C_j is reachable from class C_i. Let $\lambda^\dagger$ be the maximum over these λ_j's. If $\tilde{s}_i = 0$ and $s_i = 0$ for every inessential class C_i, then let $\lambda^\dagger = 0$.

Let $\lambda = \max\{\lambda^\star, \lambda^\dagger\}$. Then the Rényi entropy rate is given by

$$\lim_{n\to\infty} \frac{1}{n} H_\alpha(p^{(n)}) = \frac{1}{1-\alpha} \log \lambda,$$

where $0 < \alpha$, $\alpha \neq 1$.

Corollary 4.4 The rate of convergence of the Rényi entropy rate of $p^{(n)}$ is of the order $1/n$.

Although the Rényi entropy reduces to the Shannon entropy, the Rényi entropy rate does not necessarily reduce to the Shannon entropy rate as $\alpha \to 1$. From the results about the interchangeability of limits for the Rényi divergence rate as derived in Section 4.2, it follows easily that the Rényi entropy rate always reduces to the Hartley entropy rate as $\alpha \downarrow 0$ ($\lim_{n\to\infty} \frac{1}{n} H_0(p^{(n)})$), and if the Markov source is irreducible, it reduces to the Shannon entropy rate as $\alpha \to 1$. We have the following corollaries.

Corollary 4.5 Let $\alpha > 0$, $\alpha \neq 1$. Suppose that the Markov source under $p^{(n)}$ is irreducible. Then

$$\begin{aligned}\lim_{\alpha\to 1} \lim_{n\to\infty} \frac{1}{n} H_\alpha(p^{(n)}) &= \lim_{n\to\infty} \lim_{\alpha\to 1} \frac{1}{n} H_\alpha(p^{(n)}) \\ &= -\sum_{i,j} \pi_i p_{ij} \log p_{ij},\end{aligned}$$

and therefore, the Rényi entropy rate reduces to the Shannon entropy rate as $\alpha \to 1$.

Corollary 4.6 Let $\alpha > 0$, $\alpha \neq 1$. Suppose that the Markov source under $p^{(n)}$ is arbitrary (not necessarily stationary, irreducible, etc.). Then

$$\lim_{\alpha \downarrow 0} \lim_{n \to \infty} \frac{1}{n} H_\alpha(p^{(n)}) = \lim_{n \to \infty} \lim_{\alpha \downarrow 0} \frac{1}{n} H_\alpha(p^{(n)}).$$

Let us now illustrate the computation of the Rényi entropy rate by several examples. We use the natural logarithm.

Example 1: Let P be a possible probability transition matrix for $\{X_1, X_2, \ldots\}$ defined as follows:

$$P = \begin{bmatrix} 1/4 & 3/4 & 0 & 0 & 0 \\ 1/3 & 2/3 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/5 & 4/5 & 0 \\ 0 & 1/6 & 1/2 & 0 & 1/3 \end{bmatrix}.$$

Note that P is not irreducible. Indeed, P has two essential classes and 1 inessential self-communicating class. Let the parameter $\alpha = 1/3$. The largest eigenvalues of the three sub-matrices along the diagonal of R are respectively: $\lambda_1 = 1.55476$, $\lambda_2 = 1.54561$, and $\lambda_3 = 0.69336$. Let $p = (0, 0, 3/4, 1/4, 0)$ be a possible initial distribution under $p^{(n)}$. It is straightforward to check that $p^{(n)}$ is not stationary. For this given initial distribution, we get by Corollary 4.3 that $\lambda^\star = \lambda_2$ and $\lambda^\dagger = 0$. Therefore, the Rényi entropy rate is $\ln(\lambda_2)/(1-\alpha) = 0.6531$. Note that λ_2 is *not* the largest eigenvalue of R. We also obtain the following.

n	$\frac{1}{n}H_\alpha(p^{(n)})$
10	0.65368
50	0.65324
100	0.65319

Clearly, as n gets large $\frac{1}{n}H_\alpha(p^{(n)})$ is closer to the Rényi entropy rate. Note however that, in general, the function $\frac{1}{n}H_\alpha(p^{(n)})$ is not monotonic in n. Suppose that $\mathbf{s}$ has zero components on the first two classes, i.e., let $p = (0, 0, 0, 0, 1)$. In this case, $\lambda^\star = \lambda_3$, and $\lambda^\dagger = \max\{\lambda_1, \lambda_2\}$ (the first and second classes are reachable from the third). Therefore, the Rényi entropy rate is $\ln(\lambda_1)/(1 - \alpha) = 0.66198$. We also get the following.

n	$\frac{1}{n}H_\alpha(p^{(n)})$
10	0.6618
50	0.6580
100	0.6578
200	0.6582
500	0.6596

Clearly, as n gets large $\frac{1}{n}H_\alpha(p^{(n)})$ is closer to the Rényi entropy rate.

Suppose now that $\mathbf{s}$ has strictly positive components (as required in the Nemetz result). For example, let $p = (1/8, 1/4, 1/8, 1/4, 1/4)$. In this case, $\lambda^\star = \lambda^\dagger = \max\{\lambda_1, \lambda_2, \lambda_3\} = \lambda_1$. Therefore, the Rényi entropy rate is $\ln(\lambda_1)/(1 - \alpha) = 0.66198$. Note that λ_1 is the largest eigenvalue of R which is expected since the components of $\mathbf{s}$ are strictly positive. We also get the following.

n	$\frac{1}{n}H_\alpha(p^{(n)})$
10	0.7693
50	0.6800
100	0.6691

Clearly, as n gets large $\frac{1}{n}H_\alpha(p^{(n)})$ is closer to the Rényi entropy rate.

Example 2: Suppose that the Markov source is of order 2 under $p^{(n)}$ and $q^{(n)}$ respectively. Let $\{W_1, W_2, \ldots\}$ be the process obtained by 2-step blocking the Markov source. Let P be a possible transition matrix for $\{W_1, W_2, \ldots\}$ defined as follows:

$$P = \begin{bmatrix} 1/4 & 3/4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 3/5 & 2/5 & 0 & 0 \\ 0 & 0 & 1/5 & 4/5 \end{bmatrix}.$$

Note that P is not irreducible. The set of indices $\{1, 2, 3\}$ forms an essential class, while the singleton set $\{4\}$ forms a self-communicating non-essential class. Let the parameter $\alpha = 0.5$. The largest positive real eigenvalues of the two sub-matrices along the diagonal of R are respectively: $\lambda_1 = 1.24037$, $\lambda_2 = 0.89442$. Let $p = (1/4, 3/4, 0, 0)$ denote a possible initial distribution of W_1 under $p^{(n)}$. Note that $p^{(n)}$ is not stationary. For this given initial distribution, we get by Corollary 4.3 that $\lambda^* = \lambda_1$ and $\lambda^\dagger = 0$. Therefore, the Rényi entropy rate is $\ln(\lambda_1)/(1-\alpha) = 0.4308$. We also obtain the following.

n	$\frac{1}{n}H_\alpha(p^{(n)})$
10	0.3951
50	0.4236
100	0.4272

Clearly, as n gets large $\frac{1}{n}H_\alpha(p^{(n)})$ is closer to the Rényi entropy rate.

Let us now suppose that $p = (1/4, 0, 0, 3/4)$. For this given initial distribution, we get by Corollary 4.3 that $\lambda^* = \lambda_1$ and $\lambda^\dagger = \lambda_1$. Therefore, the Rényi entropy rate is $\ln(\lambda_1)/(1-\alpha) = 0.4308$. We also obtain the following.

n	$\frac{1}{n}H_\alpha(p^{(n)})$
10	0.4533
50	0.4359
100	0.4334

Clearly, as n gets large $\frac{1}{n}H_\alpha(p^{(n)})$ is closer to the Rényi entropy rate.

4.4 A Variable-Length Source Coding Theorem

Following [13], let the *average code length of order t* be defined by

$$L(t) = \frac{1}{t}\log_D\left(\sum_i p_i D^{tl_i}\right),$$

where $0 < t < \infty$, and l_i is the length of the codeword (or code sequence) for the i-th source symbol. $L(t)$ is an interesting measure of code length which implies that the cost of representing a source symbol varies *exponentially* with code length, as opposed to Shannon's expected code length measure

$$\bar{l} \triangleq \sum_{i=1}^{M} p_i l_i$$

in which the cost varies linearly with code length [13]. A simple calculation shows that $L(t)$ reduces to $\bar{l}$ when $t \to 0$; thus $L(t)$ can be regarded as a more general measure. Furthermore, in many applications where the processing cost of decoding is high or the buffer overflow due to long codewords is important, an exponential cost function can be more appropriate than a linear cost function [11], [13].

Consider a source sequence s of length n that we wish to encode via a D-ary uniquely decodable code. Let $p(s)$ be the probability of s, and $l(s)$ be the length of the codeword for s. Then the average code length of order t for the n-sequences is

$$L_n(t) = \frac{1}{t} \log_D \left(\sum_s p(s) D^{tl(s)} \right),$$

where the summation extends over the M^n sequences s. In [13], Campbell demonstrated the following variable-length source coding theorem for a DMS (discrete memoryless source), in which the Rényi entropy ($H_\alpha(p)$) plays a role *analogous* to the Shannon entropy when the cost function in the coding problem is exponential as opposed to linear.

Proposition 4.1 **[13]** Let $\alpha = 1/(1+t)$. By encoding sufficiently long sequences of input symbols of a DMS, it is possible to make the average code length of order t per input symbol $\frac{1}{n}L_n(t)$ as close to $H_\alpha(p)$ as desired. Also, it is not possible to find a uniquely decodable code whose average length of order t is less than $H_\alpha(p)$.

We next establish an operational characterization for the Rényi entropy rate by extending this theorem to Markov sources.

Theorem 4.6 Let $\alpha = 1/(1+t)$. There exists a uniquely decodable code for a Markov source with an asymptotic average code length of order t per input symbol satisfying

$$\lim_{n\to\infty} \frac{1}{n} L_n(t) = \frac{1}{1-\alpha} \log \lambda,$$

where λ denotes the positive eigenvalue of the matrix $R = (p_{ij}^{\alpha})$ as defined in Corollary 4.3. Conversely, any uniquely decodable code for the source has an asymptotic average code length of order t per input symbol satisfying

$$\lim_{n\to\infty} \frac{1}{n} L_n(t) \geq \frac{1}{1-\alpha} \log \lambda.$$

Proof: Let s be a sequence of input symbols of length n from the source. We can consider such sequence as an element from the alphabet $\mathcal{X}^M$. Proceeding exactly as in the proof of [13, Theorem 1], we can similarly establish the existence of a uniquely decodable code satisfying

$$\frac{1}{n} H_\alpha(p^{(n)}) \leq \frac{1}{n} L_n(t) < \frac{1}{n} H_\alpha(p^{(n)}) + \frac{1}{n}.$$

From Corollary 4.3, we have

$$\lim_{n\to\infty} \frac{1}{n} H_\alpha(p^{(n)}) = \frac{1}{1-\alpha} \log \lambda. \tag{4.22}$$

Therefore

$$\lim_{n\to\infty} \frac{1}{n} L_n(t) = \frac{1}{1-\alpha} \log \lambda.$$

This completes the proof of the forward part. By [13, Lemma 1], every uniquely decodable code satisfies $L_n(t) \geq H_\alpha(p^{(n)})$. Hence, the proof of the converse part follows directly from (4.22).

□

Remark: By Corollary 4.5, the above theorem does not necessarily reduce to the Shannon lossless source coding theorem as $\alpha \to 1$ and $n \to \infty$. It reduces to the Shannon coding theorem if for example the Markov source is irreducible.

Let us now illustrate numerically using a generalized Huffman code for the Markov source that $\frac{1}{n}H_\alpha(p^{(n)})$ is close to the Rényi entropy rate and that $\frac{1}{n}H_\alpha(p^{(n)})$ is close to $\frac{1}{n}L_n(t)$ for several values of n. Following [11], the Rényi redundancy of a code for a source sequence of length n is defined as

$$\rho_n = \frac{1}{n}L_n(t) - \frac{1}{n}H_\alpha(p^{(n)}).$$

In [33, Theorem 1′], a simple generalization of Huffman's algorithm which minimizes ρ_n is given. In Huffman's algorithm, each new node is assigned the weight $p_i + p_j$, where p_i and p_j are the lowest weights on available nodes. In the generalized algorithm, the new node is assigned the weight $2^t(p_i + p_j)$. The base of the logarithm is 2, so the entropies are measured in bits.

Example: Let $\{X_1, X_2, \ldots\}$ be a binary first-order Markov source with initial distribution $(0.8, 0.2)$ and probability transition matrix

$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix}.$$

Let $\alpha = 0.5$, then $t = 1$. The largest eigenvalue of $R = (p_{ij}^{\alpha})$ is found to be $\lambda = 1.396$. By Corollary 4.2, the Rényi entropy rate is equal to 0.963. Using the generalized Huffman's algorithm we get the following.

n	$\frac{1}{n}H_{\alpha}(p^{(n)})$	$\frac{1}{n}L_n(t)$
1	0.848	1.000
2	0.909	0.9705
3	0.927	0.945

The sets of codewords are (0,1), (0,10,110,111) and (10,000,001,010,110,111,0110, 0111) for $n = 1, 2$ and 3 respectively. As n gets large, $\frac{1}{n}H_{\alpha}(p^{(n)})$ is closer to the Rényi entropy rate. Also, $\frac{1}{n}L_n(t)$ is close to $\frac{1}{n}H_{\alpha}(p^{(n)})$.

Chapter 5

Csiszár's Forward Cutoff Rate for Hypothesis Testing Between General Sources with Memory

In [20], Csiszár established the concept of forward β-cutoff rate for the hypothesis testing problem based on independent and identically distributed (i.i.d.) observations. Given $\beta < 0$, he defines the forward β-cutoff rate as the number $R_0 \geq 0$ that provides the best possible lower bound in the form $\beta(E - R_0)$ to the type 1 error exponent function for hypothesis testing where $0 < E < R_0$ is the rate of exponential convergence to 0 of the type 2 error probability. He then demonstrated that the forward β-cutoff rate is given by $D_{1/(1-\beta)}(X\|\bar{X})$, where $D_\alpha(X\|\bar{X})$ denotes the α-divergence, $\alpha > 0$, $\alpha \neq 1$. This result provides a new operational significance for the α-divergence.

The error exponent for the binary hypothesis testing problem has been thoroughly studied for finite state i.i.d. sources and Markov chains. The results for i.i.d. sources can be found in [21], [31], [35], and for irreducible Markov sources in [5], [43]. The error exponent for testing between ergodic Markov sources with continuous state-space under certain additional restrictions was established in [39]. In its full generality, i.e., for arbitrary sources (not necessarily, stationary, ergodic, etc.), the error exponent was studied in [15], [29], [30].

In the sequel, we extend Csiszár's result [20] by investigating the forward β-cutoff rate for the hypothesis testing between two arbitrary sources. Our proof relies in part on the formulas established in [29], and extensions of the techniques used in [14] to generalize Csiszár's results for arbitrary discrete sources with memory. Unlike [14] where the source alphabet was assumed to be finite, we assume arbitrary (countable or continuous) source alphabet. The techniques used in our proof are a mixture of the techniques used in deriving the forward and reverse β-cutoff rates for source coding [14]. However, some new techniques are also needed to obtain the result. We demonstrate that the liminf α-divergence rate provides the expression for the forward β-cutoff rate.

5.1 Preliminaries

In this section, we briefly review previous results by Han [29] on the general expression for the Neyman-Pearson type 2 error subject to an exponential bound on the type 1 error. Let us first define the general source as an infinite sequence

$\mathbf{X} = \{X^n\}_{n=1}^{\infty} \stackrel{\triangle}{=} \left\{X^n = \left(X_1^{(n)}, \ldots, X_n^{(n)}\right)\right\}_{n=1}^{\infty}$ of n-dimensional random variables X^n where each component random variable $X_i^{(n)}$ $(1 \le i \le n)$ takes values in an arbitrary set $\mathcal{X}$ that we call the source alphabet. Given two arbitrary sources $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and $\bar{\mathbf{X}} = \{\bar{X}^n\}_{n=1}^{\infty}$ taking values in the same source alphabet $\{\mathcal{X}^n\}_{n=1}^{\infty}$, we may define the general hypothesis testing problem with $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ as the null hypothesis and $\bar{\mathbf{X}} = \{\bar{X}^n\}_{n=1}^{\infty}$ as the alternative hypothesis.

Let $\mathcal{A}_n$ be any subset of $\mathcal{X}^n$, $n = 1, 2, \ldots$ that we call the acceptance region of the hypothesis test, and define

$$\mu_n \stackrel{\triangle}{=} Pr\{X^n \notin \mathcal{A}_n\} \quad \text{and} \quad \lambda_n \stackrel{\triangle}{=} Pr\{\bar{X}^n \in \mathcal{A}_n\}$$

where μ_n, λ_n are called type 1 error probability and type 2 error probability, respectively.

Definition 5.1 Fix $r > 0$. A rate E is called r-achievable if there exists a sequence of acceptance regions $\mathcal{A}_n$ such that[1]

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mu_n \ge r \quad \text{and} \quad \liminf_{n\to\infty} -\frac{1}{n}\log \lambda_n \ge E.$$

[1] Let (a_n) be a sequence in $\mathbb{R} \cup \{-\infty, +\infty\}$. The *limit inferior* is given by

$$\begin{aligned}\liminf_{n\to\infty} a_n &= \sup_{n\ge 1} \inf_{k\ge n} a_k \\ &= \lim_{n\to\infty} \inf_{k\ge n} a_k.\end{aligned}$$

Similarly, the *limit superior* is given by

$$\begin{aligned}\limsup_{n\to\infty} a_n &= \inf_{n\ge 1} \sup_{k\ge n} a_k \\ &= \lim_{n\to\infty} \sup_{k\ge n} a_k.\end{aligned}$$

Definition 5.2 The supremum of all r-achievable rates is denoted by $B_e(r|\mathbf{X}\|\bar{\mathbf{X}})$:

$$B_e(r|\mathbf{X}\|\bar{\mathbf{X}}) \overset{\triangle}{=} \sup\{E > 0 : E \text{ is } r\text{-achievable}\},$$

and $B_e(r|\mathbf{X}\|\bar{\mathbf{X}}) = 0$ if the above set is empty.

The dual of this function is defined as:

$$D_e(E|\mathbf{X}\|\bar{\mathbf{X}}) \overset{\triangle}{=} \sup\{r > 0 : E \text{ is } r\text{-achievable}\},$$

and $D_e(E|\mathbf{X}\|\bar{\mathbf{X}}) = 0$ if the above set is empty.

Proposition 5.1 **[29]** Fix $r > 0$. For the general hypothesis testing problem, we have that

$$B_e(r|\mathbf{X}\|\bar{\mathbf{X}}) = \inf_{R\in\mathbb{R}}\{R + \eta(R) : \eta(R) < r\},$$

where[2]

$$\eta(R) \overset{\triangle}{=} \liminf_{n\to\infty} -\frac{1}{n}\log Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\bar{X}^n}(X^n)} \le R\right\},$$

is the large deviation spectrum of the normalized log-likelihood ratio.

We herein assume that the source alphabet is countable. However, we will point out the necessary modifications in the proofs for the case of a continuous alphabet. The above proposition is the main tool for our key lemma in the following section.

[2] If the source alphabet $\mathcal{X}$ is (absolutely) continuous, then $P_{X^n}(X^n)$ plays the role of the density function $f_{X^n}(X^n)$.

5.2 Hypothesis Testing Forward β-Cutoff Rate

Definition 5.3 Fix $\beta < 0$. $R_0 \geq 0$ is a forward β-achievable rate for the general hypothesis testing problem if

$$D_e(E|\mathbf{X}\|\bar{\mathbf{X}}) \geq \beta(E - R_0)$$

for every $E > 0$, or equivalently,

$$B_e(r|\mathbf{X}\|\bar{\mathbf{X}}) \geq R_0 + \frac{r}{\beta},$$

for every $r > 0$. The forward β-cutoff rate is defined as the supremum of all forward β-achievable rates, and is denoted by $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$.

Note that in the degenerate and uninteresting case where $D_e(E|\mathbf{X}\|\bar{\mathbf{X}})$ is identically 0, we have that $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) = 0$. We herein assume that $D_e(E|\mathbf{X}\|\bar{\mathbf{X}})$ is not 0 for all values of E. A graphical illustration of $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$ is presented in Figure 5.1.

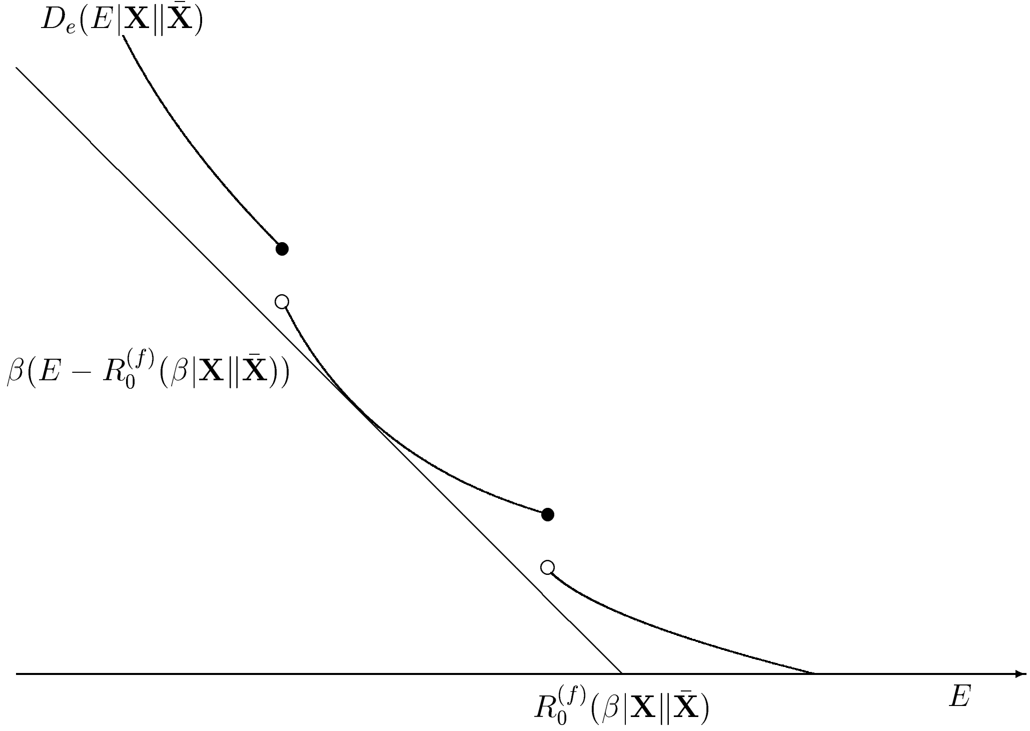
$D_e(E|\mathbf{X}\|\bar{\mathbf{X}})$
$\beta(E - R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}))$
$R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$
E

Figure 5.1: A graphical illustration of the forward β-cutoff rate, $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$, for testing between two arbitrary sources $\mathbf{X}$ and $\bar{\mathbf{X}}$.

Before stating our main result, we first observe in the next lemma that the forward β-cutoff rate $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$ is indeed the R-axis intercept of a support line of slope $\frac{\beta}{1-\beta}$ to the large deviation spectrum $\eta(R)$.

Lemma 5.1 Fix $\beta < 0$. The following conditions are equivalent.

$$(\forall R \in \mathbb{R}) \quad \eta(R) \geq \frac{\beta}{\beta - 1}(R_0 - R) \tag{5.1}$$

and

$$(\forall r > 0) \quad B_e(r|\mathbf{X}\|\bar{\mathbf{X}}) \geq R_0 + \frac{r}{\beta}. \tag{5.2}$$

Proof:

a) (5.1) ⇒ (5.2).

For any $r > 0$, we obtain by Proposition 5.1 that

$$(\forall \delta > 0)(\exists R_\delta \text{ with } \eta(R_\delta) < r) \quad B_e(r|\mathbf{X}\|\bar{\mathbf{X}}) + \delta \geq R_\delta + \eta(R_\delta).$$

Therefore

$$\begin{aligned}
B_e(r|\mathbf{X}\|\bar{\mathbf{X}}) &\geq R_\delta + \eta(R_\delta) - \delta \\
&\geq R_\delta - \delta + \frac{\beta}{\beta - 1}(R_0 - R_\delta) \qquad (5.3) \\
&= -\delta + \frac{\beta}{\beta - 1}R_0 - \frac{R_\delta}{\beta - 1} \\
&\geq -\delta + \frac{\beta}{\beta - 1}R_0 - \frac{R_0}{\beta - 1} + \frac{r}{\beta} \qquad (5.4) \\
&= \frac{r}{\beta} + R_0 - \delta,
\end{aligned}$$

where (5.3) follows from (5.1), and (5.4) holds because

$$r > \eta(R_\delta) \geq \frac{\beta}{\beta - 1}(R_0 - R_\delta).$$

Since δ can be made arbitrarily small, the proof of the forward part is completed.

b) (5.2) $\Rightarrow$ (5.1).

(5.1) holds trivially for those R satisfying $\eta(R) = \infty$. For any $R \in \mathbb{R}$ with $\eta(R) < \infty$, let $r_\delta \triangleq \eta(R) + \delta$ for some $\delta > 0$. Then (by Proposition 5.1)

$$B_e(r_\delta|\mathbf{X}\|\bar{\mathbf{X}}) \leq R + \eta(R).$$

Therefore

$$\begin{aligned} \eta(R) &\geq B_e(r_\delta|\mathbf{X}\|\bar{\mathbf{X}}) - R \\ &\geq R_0 + \frac{r_\delta}{\beta} - R \qquad (5.5) \\ &= R_0 + \frac{\eta(R)}{\beta} + \frac{\delta}{\beta} - R, \end{aligned}$$

where (5.5) follows by (5.2). Thus,

$$\eta(R) \geq \frac{\beta}{\beta - 1}(R_0 - R) + \frac{\delta}{\beta - 1}.$$

Since δ can be made arbitrarily small, the proof of the converse part is completed.

□

Theorem 5.1 (Forward $\boldsymbol{\beta}$-cutoff rate formula). Fix $\beta < 0$. For the general hypothesis testing problem,

$$R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) = \liminf_{n\to\infty} \frac{1}{n} D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n),$$

where

$$D_\alpha(X^n\|\bar{X}^n) \triangleq \frac{1}{\alpha-1}\log\left(\sum_{x^n\in\mathcal{X}^n}[P_{X^n}(x^n)]^\alpha[P_{\bar{X}^n}(x^n)]^{1-\alpha}\right)$$

is the n-dimensional α-divergence[3].

Proof: Note that $\eta(R) > 0$ for some[4] $R \in \mathbb{R}$.

1. *Forward part*: $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) \geq \liminf_{n\to\infty}\frac{1}{n}D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n)$. By the equivalence of conditions (5.1) and (5.2), it suffices to show that

$$(\forall R\in\mathbb{R})\ \eta(R) \geq \frac{\beta}{\beta-1}\left(\liminf_{n\to\infty}\frac{1}{n}D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n) - R\right).$$

Indeed, we have the following.

[3] If the source alphabet is (absolutely) continuous, i.e., it admits a density $f_{X^n}(\cdot)$, then the n-dimensional α-divergence is given by

$$D_\alpha(X^n\|\bar{X}^n) \triangleq \frac{1}{\alpha-1}\log\left(\int[f_{X^n}(x^n)]^\alpha[f_{\bar{X}^n}(x^n)]^{1-\alpha}dx^n\right).$$

[4] If $\eta(R) = 0$ for all $R\in\mathbb{R}$, then

$$B_e(r|\mathbf{X}\|\bar{\mathbf{X}}) = \inf_{R\in\mathbb{R}}\{R+\eta(R)|\eta(R)<r\} = \inf_{R\in\mathbb{R}}\{R\} = -\infty,$$

contradicting that $B_e(r|\mathbf{X}\|\bar{\mathbf{X}})$ is, by definition, an exponent and should be always non-negative.

$$\begin{aligned} Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\bar{X}^n}(X^n)} \le R\right\} &= Pr\left\{e^{-t\log\frac{P_{X^n}(X^n)}{P_{\bar{X}^n}(X^n)}} \ge e^{-ntR}\right\}, \text{ for } t>0 \\ &\le e^{ntR}\left(\sum_{x^n\in\mathcal{X}^n}[P_{X^n}(x^n)]^{1-t}[P_{\bar{X}^n}(x^n)]^t\right) \qquad (5.6) \\ &= \exp\left\{-nt\left(\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n)-R\right)\right\}, \end{aligned}$$

for $0<t<1$, where (5.6) follows by Markov's inequality. Therefore

$$\begin{aligned} \eta(R) &\ge t\left(\liminf_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n)-R\right) \\ &= \frac{\beta}{\beta-1}\left(\liminf_{n\to\infty}\frac{1}{n}D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n)-R\right), \text{ for } \beta \triangleq \frac{t}{t-1}<0. \end{aligned}$$

2. *Converse part*: $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) \le \liminf_{n\to\infty}\frac{1}{n}D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n)$.

The converse holds trivially if $\liminf_{n\to\infty}\frac{1}{n}D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n)$ is infinite. Hence we can assume that $\liminf_{n\to\infty}\frac{1}{n}D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n) < K$, where K is some constant. By the equivalence of conditions (5.1) and (5.2), it suffices to show that for any $\delta>0$ arbitrarily small, there exists $\underline{R}=\underline{R}(\delta)\in\mathbb{R}$ such that

$$\eta(\underline{R}) \le \frac{\beta}{\beta-1}\left(3\delta+\liminf_{n\to\infty}\frac{1}{n}D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n)-\underline{R}\right).$$

Consider the twisted distribution defined as:

$$\begin{aligned} P_{X^n}^{(t)}(x^n) &\triangleq \frac{[P_{\bar{X}^n}(x^n)]^t[P_{X^n}(x^n)]^{1-t}}{\sum_{\hat{x}^n\in\mathcal{X}^n}[P_{\bar{X}^n}(\hat{x}^n)]^t[P_{X^n}(\hat{x}^n)]^{1-t}} \\ &= \exp\left\{t\left[\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)}+D_{1-t}(X^n\|\bar{X}^n)\right]\right\}P_{X^n}(x^n), \qquad (5.7) \end{aligned}$$

where $t=\beta/(\beta-1)$. Note that $0<t<1$. Let $\mathcal{N}$ be a set of positive integers such that

$$\lim_{n\in\mathcal{N},n\to\infty}\frac{1}{n}D_{1/(1-\beta)}(X^n\|\bar{X}^n)=\liminf_{n\to\infty}\frac{1}{n}D_{1/(1-\beta)}(X^n\|\bar{X}^n),$$

and define

$$\tau \triangleq \sup\{R \in \mathbb{R} : \eta^{(t)}(R) > 0\},$$

where

$$\eta^{(t)}(R) \triangleq \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n}\log P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \le R\right\},$$

is the twisted large deviation spectrum of the normalized log-likelihood ratio with parameter t, and τ satisfies (cf. Lemmas 5.2 and 5.3 in Section 5.4) that

$$-\infty < \tau \le \lim_{n\in\mathcal{N}, n\to\infty} \frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) = \liminf_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) < K.$$

We then note by definition of $\eta^{(t)}(\cdot)$ and the finiteness property of τ that for any $\delta > 0$, there exists $\varepsilon > 0$ such that:

$$\eta^{(t)}(\tau-\delta) = \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n}\log P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \le \tau-\delta\right\} > \varepsilon > 0.$$

As a result,

$$P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} > \tau-\delta\right\} \ge 1 - e^{-n\varepsilon} \text{ for } n \in \mathcal{N} \text{ sufficiently large.}$$

On the other hand, define

$$\bar{\eta}^{(t)}(R) \triangleq \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n}\log P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \ge R\right\}$$

and

$$\bar{\tau} \triangleq \inf\{R \in \mathbb{R} : \bar{\eta}^{(t)}(R) > 0\}.$$

Then by noting that

$$\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} = D_{1-t}(X^n\|\bar{X}^n) - \frac{1}{t}\log\frac{P^{(t)}_{X^n}(x^n)}{P_{X^n}(x^n)},$$

we have:

$$\bar{\eta}^{(t)}(R) = \sigma\left(-tR + \frac{t}{n}D_{1-t}(X^n\|\bar{X}^n)\right)$$

and

$$\begin{aligned}\bar{\tau} &= -\frac{1}{t}\sup\{R \in \mathbb{R} : \sigma(R) > 0\} + \frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) \\ &\le \frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) \qquad (5.8) \\ &< K \quad \text{for } n \in \mathcal{N} \text{ sufficiently large,} \qquad (5.9)\end{aligned}$$

where

$$\sigma(R) \triangleq \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n}\log P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}^{(t)}(x^n)}{P_{X^n}(x^n)} \le R\right\},$$

(5.8) follows from Lemma 5.4 in Section 5.4, and (5.9) holds by definition of K. This indicates the existence of $\bar{\varepsilon} > 0$ such that $\bar{\eta}^{(t)}(K) > \bar{\varepsilon}$, which immediately gives that for $n \in \mathcal{N}$ sufficiently large,

$$P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \ge K\right\} \le e^{-n\bar{\varepsilon}}.$$

Therefore, for $n \in \mathcal{N}$ sufficiently large,

$$\begin{aligned}&P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : K > \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} > \tau - \delta\right\} \\ =\ &P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} > \tau - \delta\right\} \\ &-P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \ge K\right\} \\ \ge\ &1 - e^{-n\varepsilon} - e^{-n\bar{\varepsilon}}. \qquad (5.10)\end{aligned}$$

Let $I_1 \triangleq (\tau - \delta, b_1)$, and

$$I_k \triangleq [b_{k-1}, b_k) \quad \text{for} \quad 2 \le k \le L \triangleq \left\lceil \frac{K - \tau + \delta}{2\delta} \right\rceil,$$

where $b_k \overset{\triangle}{=} (\tau - \delta) + 2k\delta$ for $1 \leq k < L$, and $b_L \overset{\triangle}{=} K$. By (5.10), there exists $1 \leq k(n) \leq L$ such that

$$P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \in I_{k(n)}\right\} \geq \frac{1 - e^{-n\varepsilon} - e^{-n\bar{\varepsilon}}}{L}, \tag{5.11}$$

for $n \in \mathcal{N}$ sufficiently large. Then, by letting $R_1 \overset{\triangle}{=} \limsup_{n\in\mathcal{N}, n\to\infty} b_{k(n)} + \delta$, we obtain that for $n \in \mathcal{N}$ sufficiently large

$$P_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \leq R_1\right\} \geq P_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \in I_{k(n)}\right\}.$$

However, for sufficiently large $n \in \mathcal{N}$, we have that

$$\begin{aligned}
& P_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \in I_{k(n)}\right\} \\
= & \sum_{\left\{x^n\in\mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)}\in I_{k(n)}\right\}} P_{X^n}(x^n) \\
= & \sum_{\left\{x^n\in\mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)}\in I_{k(n)}\right\}} e^{-t\left(\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} + D_{1-t}(X^n\|\bar{X}^n)\right)} P^{(t)}_{X^n}(x^n) \qquad (5.12) \\
\geq & e^{-nt\left(-b_{k(n)-1} + \frac{1}{n}D_{1-t}(X^n\|\bar{X}^n)\right)} \sum_{\left\{x^n\in\mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)}\in I_{k(n)}\right\}} P^{(t)}_{X^n}(x^n) \\
= & e^{-nt\left(-b_{k(n)-1} + \frac{1}{n}D_{1-t}(X^n\|\bar{X}^n)\right)} P^{(t)}_{X^n}\left[x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \in I_{k(n)}\right] \\
\geq & \frac{1 - e^{-n\varepsilon} - e^{-n\bar{\varepsilon}}}{L} e^{-nt\left(-b_{k(n)-1} + \frac{1}{n}D_{1-t}(X^n\|\bar{X}^n)\right)}, \qquad (5.13)
\end{aligned}$$

where (5.12) follows from (5.7), and (5.13) follows from (5.11). Consequently

$$
\begin{aligned}
\eta(R_1) &= \liminf_{n\to\infty} -\frac{1}{n}\log P_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \le R_1\right\} \\
&\le \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n}\log P_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \le R_1\right\} \\
&\le t\left(-\limsup_{n\in\mathcal{N}, n\to\infty} b_{k(n)-1} + \liminf_{n\in\mathcal{N}, n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n)\right) \\
&\le t\left(-\limsup_{n\in\mathcal{N}, n\to\infty} b_{k(n)} + 2\delta + \liminf_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n)\right) \\
&= t\left(3\delta + \liminf_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) - R_1\right).
\end{aligned}
$$

Since δ can be made arbitrarily small, the proof is completed. □

Observations:

A. While the proof of the forward part is straightforward, the proof of the converse part is considerably more complex. The objective of the converse part is to demonstrate that if $\liminf_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n)$ is slightly shifted to the right (by a factor of 3δ), then there exists a coordinate $\underline{R}$ such that a straight line of slope $\beta/(1-\beta)$ given by

$$
y = \frac{\beta}{\beta-1}\left(3\delta + \liminf_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) - R\right)
$$

lies above the curve of $\eta(R)$ at $R = \underline{R}$, thus violating its status of support line for $\eta(R)$.

This proof is established by observing that the desired coordinate $\underline{R}$ lies in a small neighborhood of τ, where τ is the smallest point for which $\eta^{(t)}(R)$ vanishes. A key point is to choose the twisted parameter t to be equal to $\beta/(\beta-1)$ which is the negative slope of the support line to $\eta(R)$. We graphically illustrate this observation (based on a true example) in Figure 5.2. The computational details are described in Example 1 (cf. Section 5.3).

B. Note also that the proof holds if the alphabet is *countable* or *continuous* as opposed to the source coding cutoff rate [14] where the finiteness property of the alphabet is necessary. The modifications in the proof for the continuous case are clear. Simply, replace the probability mass function by the probability density function and the summation by integration. We graphically illustrate this observation (based on a true example involving memoryless Gaussian sources) in Figure 5.3. The computational details are described in Example 2 (cf. Section 5.3).

C. The proof of the hypothesis testing cutoff rate is more involved than the proof of the source coding cutoff rate given in [14]. The main difficulty arises from the formula in Theorem 5.1 where the infimum for R is taken over the entire real line contrary to Theorem 1 in [14] for source coding where R ranges from 0 to ∞. This requires us to deal separately with the degenerate case $\tau=-\infty$ (cf. Lemma 5.3 in Section 5.4). Also, the technique used to prove the forward cutoff rate for hypothesis testing relies on the proofs of *both* the source coding forward and reverse cutoff rates, but in major parts though similar to the reverse source coding cutoff rate.

D. If the sources $\mathbf{X}$ and $\bar{\mathbf{X}}$ are arbitrary (not necessarily stationary, irreducible) Markov sources of arbitrary order, then we know from Chapter 4 that the α-divergence rate exists and can be computed. Thus in this case, the forward β-cutoff rate for testing between Markov sources can be obtained. Also, from the definition of $D_e(E|\mathbf{X}\|\bar{\mathbf{X}})$, it follows directly that for all $E > 0$,

$$D_e(E|\mathbf{X}\|\bar{\mathbf{X}}) \geq \sup_{\beta<0} \left[\beta(E - R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}))\right].$$

Note that this convex lower bound is *computable* for the entire class of Markov sources, while $D_e(E|\mathbf{X}\|\bar{\mathbf{X}})$ is not necessarily computable in general (it is computable for irreducible Markov sources [5], [43], see Figure 5.4). We graphically illustrate this observation for testing between irreducible Markov sources in Figure 5.4 and arbitrary Markov sources (not necessarily stationary, irreducible) in Figure 5.5. The computational details are described in Examples 3 and 4 (cf. Section 5.3).

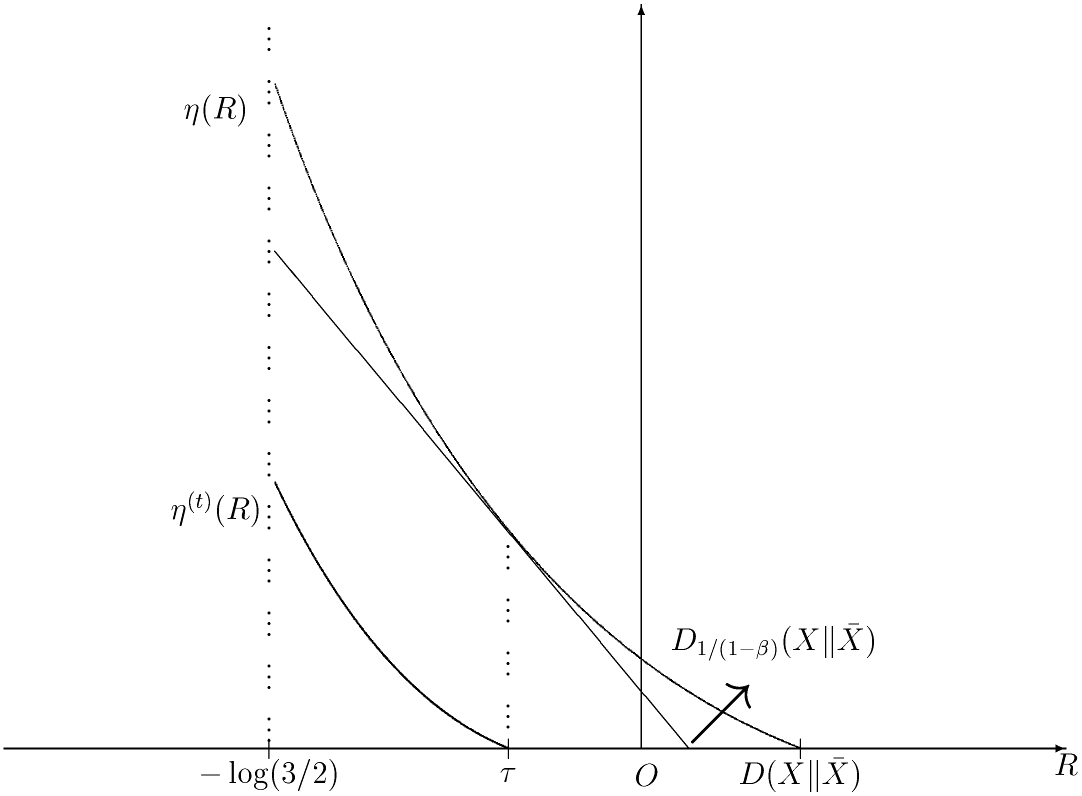

Figure 5.2: Functions $\eta(R)$, $\eta^{(t)}(R)$ and $(\beta/(\beta-1))\left[\liminf_{n\to\infty}\frac{1}{n}D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n)-R\right]$ for testing between two binary memoryless sources $\mathbf{X}=\{X_i\}_{i=1}^{\infty}$ and $\bar{\mathbf{X}}=\{\bar{X}_i\}_{i=1}^{\infty}$ under the distributions $(1/2,1/2)$ and $(1/4,3/4)$ respectively, and with $\beta=-7$. When $R<-\log(3/2)$, $\eta(R)=\eta^{(t)}(R)=\infty$.

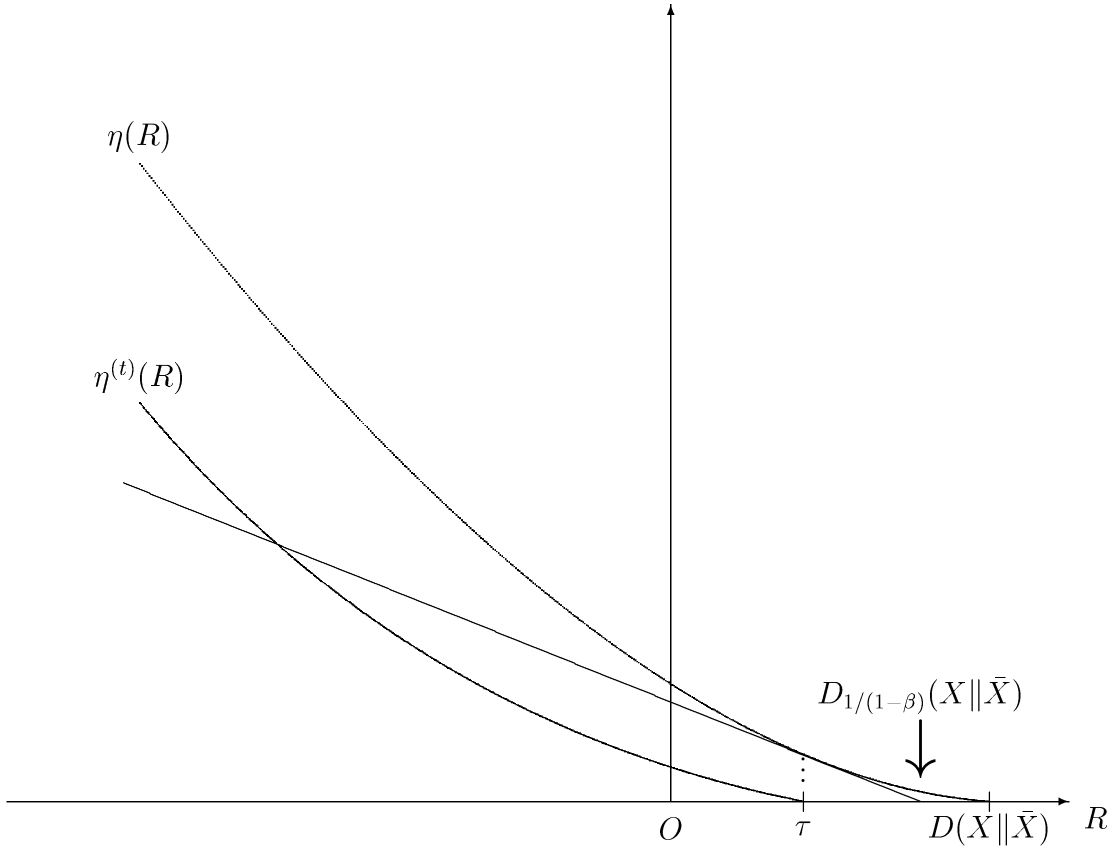

Figure 5.3: Functions $\eta(R)$, $\eta^{(t)}(R)$ and $(\beta/(\beta-1))\left[\liminf_{n\to\infty}\frac{1}{n}D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n)-R\right]$ for testing between two memoryless sources $\mathbf{X}=\{X_i\}_{i=1}^{\infty}$ and $\bar{\mathbf{X}}=\{\bar{X}_i\}_{i=1}^{\infty}$ under the Gaussian distributions $N(\nu,1)$ and $N(-\nu,1)$ respectively, and with $\beta=-0.5$.

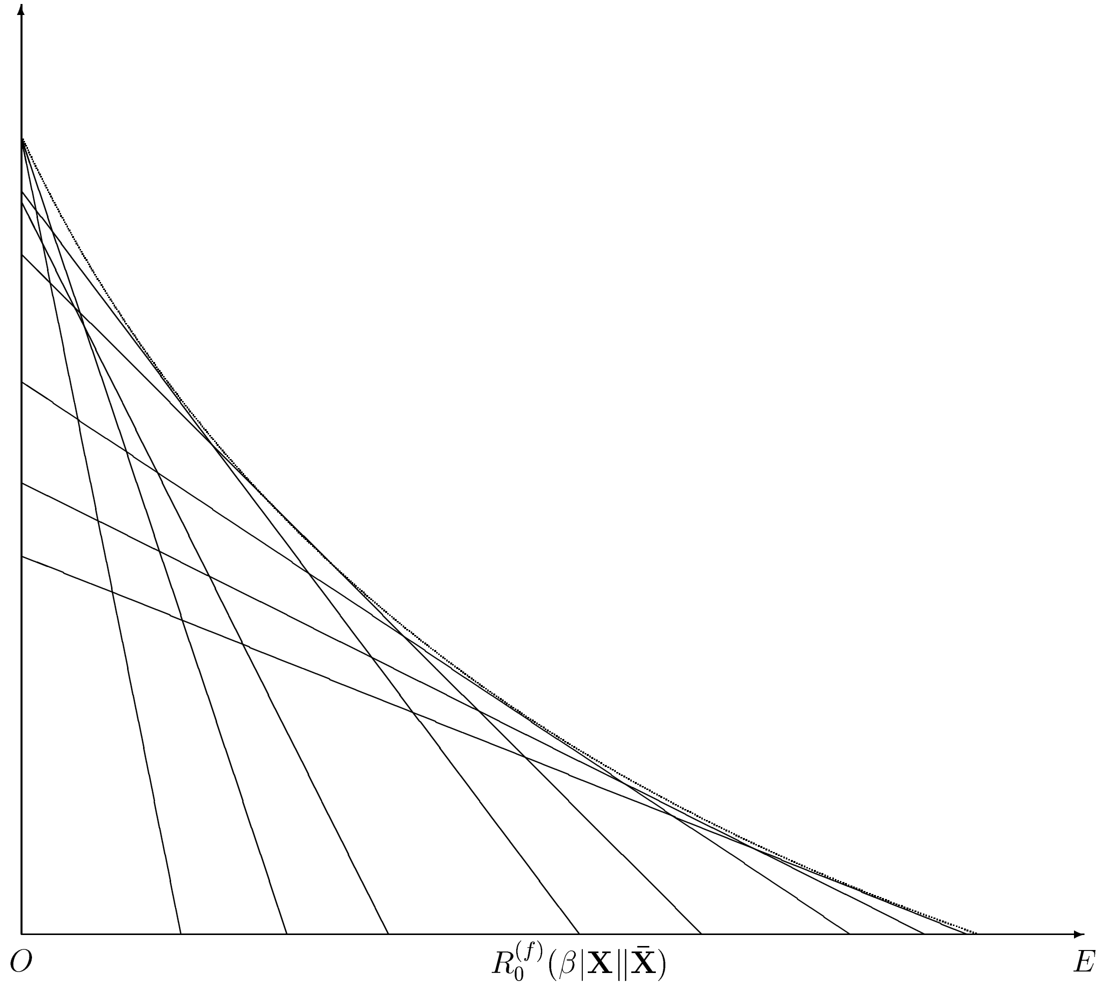

Figure 5.4: Convex lower bound for testing between irreducible Markov sources. Each line of slope β intersects the E-axis at $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$. Proceeding from left to right, the values of β are: $-5, -3, -2, -4/3, -1, -2/3, -1/2, -2/5$.

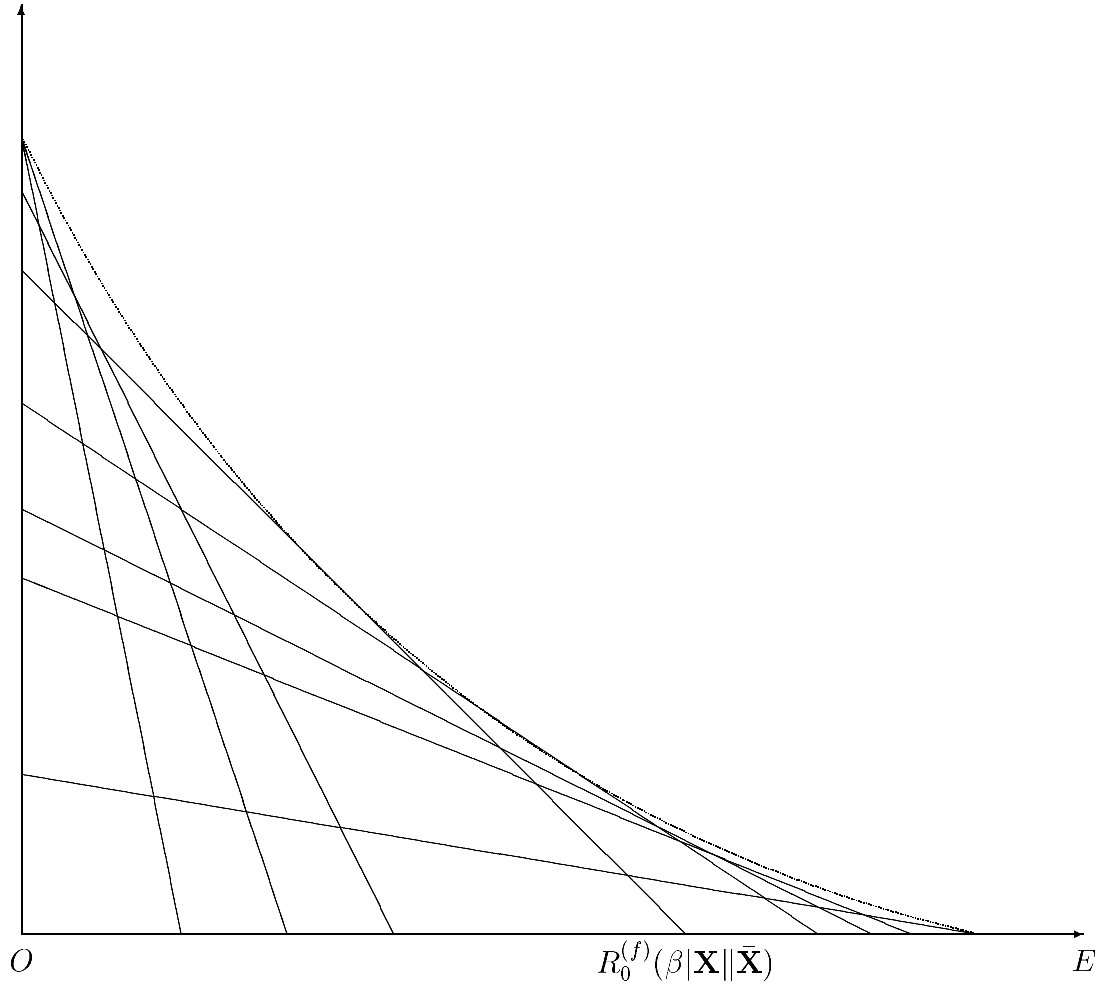

Figure 5.5: Convex lower bound for testing between arbitrary Markov sources. Each line of slope β intersects the E-axis at $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$. Proceeding from left to right, the values of β are: $-5, -3, -2, -1, -2/3, -1/2, -2/5, -1/6$.

5.3 Numerical Examples

Throughout this section, the natural logarithm is used.

Example 1 *Finite-alphabet memoryless sources:* Consider the binary hypothesis testing between two memoryless sources $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ and $\bar{\mathbf{X}} = \{\bar{X}_i\}_{i=1}^{\infty}$ under the distributions $(1/2, 1/2)$ and $(1/4, 3/4)$ respectively. Then the log-likelihood ratio $Z = \log \frac{P_X(X)}{P_{\bar{X}}(X)}$ has the following distribution:

$$Pr\{Z = \log(2)\} = 1 - Pr\{Z = \log(2/3)\} = 1/2.$$

Let $M_Z(\theta)$ denote the moment generating function of the random variable Z. By Cramer's Theorem[5] [12, p. 9], we get that

[5] Let $\{Y_1, Y_2, \ldots\}$ be an i.i.d. sequence of random variables. Suppose that the expected value of Y_1, $E[Y_1]$, exists and is finite. Consider the function

$$I(y) \triangleq \sup_{\theta \in \mathbb{R}} [\theta y - \log M(\theta)],$$

where $M(\theta) \triangleq E\{\exp[\theta Y_1]\}$ is the moment generating function of Y_1. The function $I(y)$ is known as the large deviation rate function. A simple version of Cramer's Theorem is as follows. Assume that $M(\theta) < \infty$ for all θ. For $a \geq E[Y_1]$,

$$\liminf_{n \to \infty} -\frac{1}{n} \log Pr\{S_n \leq a\} = \limsup_{n \to \infty} -\frac{1}{n} \log Pr\{S_n \leq a\} = 0$$

where $S_n \triangleq \frac{Y_1 + \cdots + Y_n}{n}$ is the sample average. This follows directly from the law of large numbers. For $a < E[Y_1]$,

$$\liminf_{n \to \infty} -\frac{1}{n} \log Pr\{S_n \leq a\} = \limsup_{n \to \infty} -\frac{1}{n} \log Pr\{S_n \leq a\} = I(a).$$

$$
\begin{aligned}
\eta(R) &= \inf_{\lambda \in (-\infty, R]} I_Z(\lambda) \\
&= \begin{cases} I_Z(R), & R < E[Z] = \log(2) - \log(3)/2; \\ 0, & \text{otherwise,} \end{cases}
\end{aligned}
$$

where $E[Z]$ denotes the expectation of the random variable Z and

$$
\begin{aligned}
I_Z(\lambda) &= \sup_{\theta \in \mathbb{R}} (\theta\lambda - \log M_Z(\theta)) \\
&= \sup_{\theta \in \mathbb{R}} \left(\theta\lambda - (\theta - 1)\log(2) - \log(1 + 3^{-\theta})\right) \\
&= \frac{\log(\log(3/2) + \lambda) - \log(\log(2) - \lambda)}{\log(3)}(\lambda - \log(2)) + \log(2) \\
&\quad - \log(1 + \frac{\log(2) - \lambda}{\log(3/2) + \lambda}) \\
&= \frac{\log(\log(3/2) + \lambda) - \log(\log(2) - \lambda)}{\log(3)}\lambda + \frac{\log(3/2)}{\log(3)}\log(\log(3/2) + \lambda) \\
&\quad + \frac{\log(2)}{\log(3)}\log(\log(2) - \lambda) + \log(2) - \log(\log(3)).
\end{aligned}
$$

Consequently,

$$
\eta(R) = \begin{cases}
\infty, & R < -\log(3/2) \\
\log(2), & R = -\log(3/2) \\
\frac{\log(\log(3/2)+R) - \log(\log(2)-R)}{\log(3)} R & \\
+ \frac{\log(3/2)}{\log(3)} \log(\log(3/2) + R) & \\
+ \frac{\log(2)}{\log(3)} \log(\log(2) - R) & \\
+ \log(2) - \log(\log(3)), & -\log(3/2) < R < \log(2) - \log(3)/2 \\
0, & \text{otherwise.}
\end{cases}
$$

Let R' be the rate at which the line of slope $\beta/(1-\beta)$ is tangent to $\eta(R)$. We have that $\eta'(R)|_{R=R'} = \beta/(1-\beta)$. Note that

$$\begin{aligned}
\eta'(R) &= \frac{R}{\log 3}\left(\frac{1}{R+\log(3/2)}+\frac{1}{\log 2 - R}\right)+\frac{\log(3/2)}{\log 3}\frac{1}{R+\log(3/2)} \\
&\quad +\frac{1}{\log 3}\left(\log(\log(3/2)+R)-\log(\log 2 - R)\right)-\frac{\log 2}{\log 3}\frac{1}{\log 2 - R} \\
&= \frac{1}{\log 3}\frac{R+\log(3/2)}{\log 2 - R}.
\end{aligned}$$

Hence

$$\frac{1}{\log 3}\frac{R'+\log(3/2)}{\log 2 - R'} = \frac{\beta}{1-\beta},$$

which yields

$$R' = \log 2 - \log\frac{3}{1+3^{\frac{\beta}{1-\beta}}}.$$

By straightforward calculations we get that

$$\eta(R') = \left(1-\frac{1}{1+3^{\frac{\beta}{1-\beta}}}\right)\log 3^{\frac{\beta}{1-\beta}} + \log 2 - \log\left(1+3^{\frac{\beta}{1-\beta}}\right).$$

Thus, the forward cutoff rate $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$, which is the R-axis intercept of the line of slope $\beta/(1-\beta)$, is given by

$$\begin{aligned}
R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) &= \frac{\beta-1}{\beta}\eta(R') + R' \\
&= \frac{2\beta-1}{\beta}\log 2 - \frac{\beta-1}{\beta}\log\left(1+3^{\frac{\beta}{1-\beta}}\right) - \log 3.
\end{aligned}$$

On the other hand, the α-divergence between $\mathbf{X}$ and $\bar{\mathbf{X}}$ is given by

$$\begin{aligned}
D_\alpha(X\|\bar{X}) &= \frac{1}{\alpha-1}\log\left(\left(\frac{1}{2}\right)^{\alpha}\left(\frac{1}{4}\right)^{1-\alpha}+\left(\frac{1}{2}\right)^{\alpha}\left(\frac{3}{4}\right)^{1-\alpha}\right) \\
&= \frac{1}{\alpha-1}\left((\alpha-2)\log 2 + \log(1+3^{1-\alpha})\right),
\end{aligned}$$

which yields

$$D_{\frac{1}{1-\beta}}(X\|\bar{X}) = \frac{2\beta-1}{\beta}\log 2 - \frac{\beta-1}{\beta}\log\left(1+3^{\frac{\beta}{1-\beta}}\right) - \log 3.$$

Note that the forward cutoff rate $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$ and the lim inf α-divergence rate (which is equal to the α-divergence since the sources are DMS) of order $\alpha = 1/(1-\beta)$ are equal as expected from Theorem 5.1. Let us now derive τ in order to check that $\tau = R'$. First, we need to compute $\eta^{(t)}(R)$. The set $\mathcal{N}$ is equal to the set of natural numbers in this case. Note that the distribution of the random variable $Z^{(t)}$ under the twisted distribution with parameter $0 < t < 1$ is given by

$$P^{(t)}\{Z = \log 2\} = 1/(1+3^t) \quad \text{and} \quad P^{(t)}\{Z = \log(2/3)\} = 3^t/(1+3^t).$$

By Cramer's theorem [12, p. 9], we get that

$$\begin{aligned}\eta^{(t)}(R) &= \inf_{\lambda\in(-\infty,R]} I_{Z^{(t)}}(\lambda)\\ &= \begin{cases} I_Z^{(t)}(R), & R < E_{P^{(t)}}[Z^{(t)}] = \frac{\log 2}{1+3^t} + \log(2/3)\frac{3^t}{1+3^t};\\ 0, & \text{otherwise,}\end{cases}\end{aligned}$$

where $E_{P^{(t)}}[Z^{(t)}]$ denotes the expectation of the random variable $Z^{(t)}$ under the twisted distribution and

$$\begin{aligned}I_Z^{(t)}(\lambda) &= \sup_{\theta\in\mathbb{R}}\left(\theta\lambda - \log M_Z^{(t)}(\theta)\right)\\ &= \sup_{\theta\in\mathbb{R}}\left(\theta\lambda - \theta\log(2) - \log(1+3^{t-\theta}) + \log(1+3^t)\right)\\ &= \left\{t + \frac{1}{\log 3}\left[\log(\lambda+\log(3/2)) - \log(\log 2 - \lambda)\right]\right\}(\lambda - \log 2)\\ &\quad + \log(1+3^t) - \log\left(1 + \frac{\log 2 - \lambda}{\lambda - \log 2 + \log 3}\right).\end{aligned}$$

Finally, we get that

$$\eta^{(t)}(R) = \begin{cases} \infty, & R < -\log(3/2) \\ \log(1+3^t), & R = -\log(3/2) \\ t(R - \log 2) & \\ +\frac{\log(\log(3/2)+R) - \log(\log(2)-R)}{\log(3)} R & \\ +\frac{\log(3/2)}{\log(3)} \log(\log(3/2)+R) & \\ +\frac{\log(2)}{\log(3)} \log(\log(2)-R) & \\ +\log(1+3^t) - \log(\log(3)), & -\log(3/2) < R < \frac{\log 2}{1+3^t} + \log(2/3)\frac{3^t}{1+3^t} \\ 0, & \text{otherwise.} \end{cases}$$

Therefore

$$\tau = \frac{\log 2}{1+3^t} + \log(2/3)\frac{3^t}{1+3^t}.$$

It is easy to check that indeed we have $\tau = R'$ when the twisted parameter t is chosen to be $\beta/(\beta-1)$. This example is illustrated in Figure 5.2 for $\beta = -7$.

Example 2 *Continuous alphabet memoryless sources:* Consider the hypothesis testing problem between two memoryless sources $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ and $\bar{\mathbf{X}} = \{\bar{X}_i\}_{i=1}^{\infty}$ under the Gaussian distributions $N(\nu, 1)$ and $N(-\nu, 1)$ respectively. It is easy to check that the log-likelihood ratio Z is Gaussian distributed with mean $2\nu^2$ and variance $4\nu^2$, which gives that the moment generating function of Z is $E[e^{\theta Z}] = e^{2\nu^2\theta + 2\nu^2\theta^2}$. So, $I_Z(\lambda) = \sup_{\theta\in\mathbb{R}}(\theta\lambda - 2\nu^2\theta - 2\nu^2\theta^2) = (\lambda - 2\nu^2)^2/(8\nu^2)$. By Cramer's theorem, we get that

$$\eta(R) = \begin{cases} \frac{1}{8\nu^2}(R - 2\nu^2)^2, & R < 2\nu^2 \\ 0, & \text{otherwise.} \end{cases}$$

Let R' be the rate at which the line of slope $\beta/(1-\beta)$ is tangent to $\eta(R)$. We have that $\eta'(R)|_{R=R'} = \beta/(1-\beta)$. Note that

$$\eta'(R) = \frac{1}{4\nu^2}(R - 2\nu^2).$$

Hence

$$\frac{1}{4\nu^2}(R' - 2\nu^2) = \frac{\beta}{\beta - 1},$$

which yields

$$R' = 2\nu^2 \frac{1+\beta}{1-\beta}\cdot$$

By straightforward calculations we get that

$$\eta(R') = \frac{2\nu^2\beta^2}{(1-\beta)^2}\cdot$$

Thus, the forward cutoff rate $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$, which is the R-axis intercept of the line of slope $\beta/(1-\beta)$, is given by

$$\begin{aligned} R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) &= \frac{\beta - 1}{\beta}\eta(R') + R' \\ &= 2\nu^2 \frac{1}{1-\beta}\cdot \end{aligned}$$

On the other hand, the α-divergence between $\mathbf{X}$ and $\bar{\mathbf{X}}$ is given by

$$\begin{aligned} D_\alpha(X\|\bar{X}) &= \frac{1}{\alpha-1}\log\int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\alpha(x-\nu)^2 - \frac{1}{2}(1-\alpha)(x+\nu)^2} dx \\ &= \frac{1}{\alpha-1}\log e^{-\frac{1}{2}(\nu^2-(2\alpha\nu-\nu)^2)} \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-(2\alpha\nu-\nu)^2)} dx \\ &= \frac{1}{\alpha-1}\log e^{-\frac{1}{2}(\nu^2-(2\alpha\nu-\nu)^2)} \\ &= 2\nu^2\alpha \end{aligned}$$

which yields

$$D_{\frac{1}{1-\beta}}(X\|\bar{X}) = 2\nu^2 \frac{1}{1-\beta}.$$

Note that the forward cutoff rate $R_0^{(f)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$ and the lim inf α-divergence rate (which is equal to the α-divergence since the sources are DMS) of order $\alpha = 1/(1-\beta)$ are equal as expected from Theorem 5.1.

Now, let us compute $\eta^{(t)}(R)$. The set $\mathcal{N}$ in this case is equal to the set of natural numbers. For some normalization constant C,

$$\begin{aligned} P_{X^n}^{(t)}(x^n) &= C \cdot \exp\left\{-\frac{t}{2}\sum_{i=1}^{n}(x_i+\nu)^2\right\}\exp\left\{-\frac{1-t}{2}\sum_{i=1}^{n}(x_i-\nu)^2\right\} \\ &= C \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}[t(x_i+\nu)^2+(1-t)(x_i-\nu)^2]\right\} \\ &= C \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i^2+2(2t-1)\nu x_i+\nu^2)\right\}, \end{aligned}$$

which is a Gaussian distribution with mean $(1-2t)\nu$ and unit variance. Similarly, by invoking Cramer's theorem, we get that,

$$\eta^{(t)}(R) = \begin{cases} \frac{1}{8\nu^2}(R+(2t-1)2\nu^2)^2, & R < (1-2t)2\nu^2 \\ 0, & \text{otherwise.} \end{cases}$$

Hence, $\tau = (1-2t)2\nu^2$. It is straightforward to check that $\tau = R'$ when the twisted parameter t is chosen to be $\beta/(\beta-1)$. This example is depicted in Figure 5.3 for $\beta = -0.5$.

Example 3 *Irreducible finite-alphabet Markov sources:* Suppose that $\mathbf{X}$ and $\bar{\mathbf{X}}$ are two irreducible Markov sources with arbitrary initial distributions and probability transition matrices P and Q defined as follows:

$$P = \begin{pmatrix} 1/3 & 2/3 \\ 1/4 & 3/4 \end{pmatrix}, \qquad Q = \begin{pmatrix} 1/5 & 4/5 \\ 5/6 & 1/6 \end{pmatrix}.$$

Define a new matrix $R = (r_{ij})$ by

$$r_{ij} = p_{ij}^{\alpha} q_{ij}^{1-\alpha}, \quad i, j = 0, 1.$$

By Theorem 4.1, the α-divergence rate between $\mathbf{X}$ and $\bar{\mathbf{X}}$ exists and is given by

$$\lim_{n\to\infty} \frac{1}{n} D_\alpha(X^n \| \bar{X}^n) = \frac{1}{\alpha - 1} \log \lambda,$$

where λ is the largest positive real eigenvalue of R. Hence the computation of the convex lower bound for $D_e(E|\mathbf{X}\|\bar{\mathbf{X}})$ is easily obtained as shown in Figure 5.4 for the values $\beta = -5, -3, -2, -4/3, -1,\ -2/3, -1/2, -2/5$ (proceeding from left to right), where $\alpha = \frac{1}{1-\beta}$. Note that in this case the bound is tight [5], [43].

Example 4 *Arbitrary finite-alphabet Markov sources:* Suppose that $\mathbf{X}$ and $\bar{\mathbf{X}}$ are two arbitrary Markov sources with arbitrary initial distributions and probability transition matrices P and Q defined as follows:

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 & 0 \\ 0 & 0 & 3/5 & 2/5 & 0 \\ 0 & 1/6 & 5/6 & 0 & 0 \\ 1/4 & 0 & 1/4 & 0 & 1/2 \end{pmatrix}, \qquad Q = \begin{pmatrix} 1/5 & 4/5 & 0 & 0 & 0 \\ 2/3 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1/6 & 5/6 & 0 & 0 \\ 1/8 & 0 & 1/2 & 0 & 3/8 \end{pmatrix}.$$

Define a new matrix $R = (r_{ij})$ by

$$r_{ij} = p_{ij}^{\alpha} q_{ij}^{1-\alpha}, \quad i, j = 0, 1, 2, 3, 4.$$

By Theorem 4.2, the α-divergence rate between $\mathbf{X}$ and $\bar{\mathbf{X}}$ can be computed. Hence, the convex lower bound for $D_e(E|\mathbf{X}\|\bar{\mathbf{X}})$ can be easily derived as shown in Figure 5.5 for the values $\beta = -5, -3, -2, -1, -2/3, -1/2, -2/5, -1/6$ (proceeding from left to right), where $\alpha = \frac{1}{1-\beta}$.

5.4 Properties of τ and $\sigma(R)$

Lemma 5.2 For $0 < t < 1$,

$$\tau \triangleq \sup\{R : \eta^{(t)}(R) > 0\} \leq \liminf_{n\to\infty} \frac{1}{n} D_{1-t}(X^n \| \bar{X}^n).$$

Proof: For any $\nu > 0$,

$$\begin{aligned} & P_{X^n}^{(t)} \left\{ x^n \in \mathcal{X}^n : \frac{1}{n} \log \frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} > \liminf_{n\to\infty} \frac{1}{n} D_{1-t}(X^n \| \bar{X}^n) + 2\nu \right\} \\ \leq \; & P_{X^n}^{(t)} \left\{ x^n \in \mathcal{X}^n : \frac{1}{n} \log \frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} > \frac{1}{n} D_{1-t}(X^n \| \bar{X}^n) + \nu \right\} \end{aligned}$$

for sufficiently large $n \in \mathcal{N}$. But

$$\begin{aligned}
& P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} > \frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) + \nu\right\} \\
= \;& P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : -\frac{1}{n}\left(\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} + D_{1-t}(X^n\|\bar{X}^n)\right) > \nu\right\} \\
= \;& P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{t}{n}\left(\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} + D_{1-t}(X^n\|\bar{X}^n)\right) < -\nu t\right\} \\
= \;& P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P^{(t)}_{X^n}(x^n)}{P_{X^n}(x^n)} < -\nu t\right\} \qquad (5.14) \\
= \;& P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : P^{(t)}_{X^n}(x^n) < e^{-n\nu t}P_{X^n}(x^n)\right\} \\
\le \;& e^{-n\nu t}P_{X^n}\left\{x^n \in \mathcal{X}^n : P^{(t)}_{X^n}(x^n) < e^{-n\nu t}P_{X^n}(x^n)\right\} \\
\le \;& e^{-n\nu t},
\end{aligned}$$

where (5.14) follows from (5.7). Thus for sufficiently large $n \in \mathcal{N}$,

$$P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \le \liminf_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) + 2\nu\right\} \ge 1 - e^{-n\nu t},$$

which implies

$$\begin{aligned}
& \eta^{(t)}\left(\liminf_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) + 2\nu\right) \\
= \;& \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n}\log P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \le \liminf_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) + 2\nu\right\} \\
\le \;& \limsup_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n}\log\left(1 - e^{-n\nu t}\right) = 0.
\end{aligned}$$

Consequently,

$$\sup\left\{R : \eta^{(t)}(R) > 0\right\} \le \liminf_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) + 2\nu.$$

The proof is completed by noting that ν can be made arbitrarily small.

□

Lemma 5.3 For $0 < t < 1$, if $\liminf_{n\to\infty} \frac{1}{n} D_{1-t}(X^n \| \bar{X}^n) < K$, then

$$\tau \triangleq \sup\{R : \eta^{(t)}(R) > 0\} > -\infty.$$

Proof: By (5.7), we get that

$$P^{(t)}_{X^n}(x^n) \;\; = \;\; e^{tD_{1-t}(X^n\|\bar{X}^n)} e^{(1-t)\log \frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)}} P_{\bar{X}^n}(x^n).$$

Hence,

$$\begin{aligned} & P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \le R\right\} \\ \le \;\; & e^{tD_{1-t}(X^n\|\bar{X}^n)} e^{(1-t)nR} P_{\bar{X}^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \le R\right\} \\ \le \;\; & e^{tD_{1-t}(X^n\|\bar{X}^n)} e^{(1-t)nR}, \end{aligned}$$

which implies that

$$\eta^{(t)}(R) \ge -t \limsup_{n\in\mathcal{N}, n\to\infty} \frac{1}{n} D_{1-t}(X^n\|\bar{X}^n) - (1-t)R.$$

Therefore,

$$\tau \ge -\frac{t}{1-t} \limsup_{n\in\mathcal{N}, n\to\infty} \frac{1}{n} D_{1-t}(X^n\|\bar{X}^n).$$

This shows that $\tau = -\infty$ implies that

$$\limsup_{n\in\mathcal{N}, n\to\infty} \frac{1}{n} D_{1-t}(X^n\|\bar{X}^n) = \lim_{n\in\mathcal{N}, n\to\infty} \frac{1}{n} D_{1-t}(X^n\|\bar{X}^n) = \liminf_{n\to\infty} \frac{1}{n} D_{1-t}(X^n\|\bar{X}^n) = \infty,$$

contradicting the assumption that $\liminf_{n\to\infty}(1/n)D_{1-t}(X^n\|\bar{X}^n) < K$.

$\square$

Lemma 5.4 We have the following:

$$\sup\{R \in \mathbb{R} : \sigma(R) > 0\} \geq 0.$$

Proof: For any $\nu > 0$,

$$\begin{aligned}
& P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}^{(t)}(x^n)}{P_{X^n}(x^n)} \leq -\nu\right\} \\
= \; & P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : P_{X^n}^{(t)}(x^n) \leq e^{-n\nu}P_{X^n}(x^n)\right\} \\
\leq \; & e^{-n\nu}P_{X^n}\left\{x^n \in \mathcal{X}^n : P_{X^n}^{(t)}(x^n) \leq e^{-n\nu}P_{X^n}(x^n)\right\} \\
\leq \; & e^{-n\nu},
\end{aligned}$$

which implies $\sigma(-\nu) \geq \nu$. Hence, the lemma holds. □

Chapter 6

Csiszár's Reverse Cutoff Rate for Hypothesis Testing Between General Sources with Memory

In [20], Csiszár established the concept of reverse β-cutoff rate for the hypothesis testing problem based on i.i.d. observations. Given $\beta > 0$, he defines the reverse β-cutoff rate as the number $R_0 \geq 0$ that provides the best possible lower bound in the form $\beta(E - R_0)$ to the type 1 correct exponent (or reliability) function for hypothesis testing where $0 < R_0 < E$ is the rate of exponential convergence to 0 of the type 2 error probability. He then demonstrated that the reverse β-cutoff rate is given by $D_{1/(1-\beta)}(X\|\bar{X})$, where $D_\alpha(X\|\bar{X})$ denotes the α-divergence, $\alpha > 0$, $\alpha \neq 1$. This result provides a new operational significance for the α-divergence.

In this chapter, we extend Csiszár's result [20] by investigating the reverse β-

cutoff rate for hypothesis testing between two arbitrary sources. Our proof relies in part on the formulas established in [29], and extensions of the techniques used in [14] to generalize Csiszár's source coding result for arbitrary discrete sources. Unlike [14] where the source alphabet was assumed to be finite, we assume arbitrary (countable or continuous) source alphabet. We show that if the log-likelihood ratio large deviation spectrum $\rho(R)$ is convex and if there exists an $R \in \mathbb{R}$ such that $\rho(R) + R = 0$, then the limsup α-divergence rate with $\alpha = \frac{1}{1-\beta}$ provides the expression for the reverse β-cutoff rate for $0 < \beta < \beta_{\max}$, where $\beta_{\max}$ is the largest $\beta < 1$ for which the $\limsup$ $\frac{1}{1-\beta}$-divergence rate is finite. For $1 > \beta \geq \beta_{\max}$, we only provide an upper bound on the reverse cutoff rate. However, our result does reduce to Csiszár's result for finite-alphabet i.i.d. observations for $0 < \beta < 1$. In the following section, relevant previous results by Han on the probability of correct testing are briefly reviewed and the problem setup is presented.

6.1 Preliminaries and Problem Formulation

Define the general source [29] as an infinite sequence $\mathbf{X} = \{X^n\}_{n=1}^{\infty} \triangleq \left\{X^n = \left(X_1^{(n)}, \ldots, X_n^{(n)}\right)\right\}_{n=1}^{\infty}$ of n-dimensional random variables X^n where each component random variable $X_i^{(n)}$ $(1 \leq i \leq n)$ takes values in an arbitrary set $\mathcal{X}$ that we call the source alphabet. Given two arbitrary sources $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and $\bar{\mathbf{X}} = \{\bar{X}^n\}_{n=1}^{\infty}$ taking values in the same source alphabet $\{\mathcal{X}^n\}_{n=1}^{\infty}$, we may define the general hypothesis testing problem with $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ as the null hypothesis and $\bar{\mathbf{X}} = \{\bar{X}^n\}_{n=1}^{\infty}$ as the alternative hypothesis.

Let $\mathcal{A}_n$ be any subset of $\mathcal{X}^n$, $n = 1, 2, \ldots$ that we call the acceptance region of the hypothesis test, and define

$$\mu_n \triangleq Pr\{X^n \notin \mathcal{A}_n\} \quad \text{and} \quad \lambda_n \triangleq Pr\{\bar{X}^n \in \mathcal{A}_n\}$$

where μ_n, λ_n are called type 1 error probability and type 2 error probability, respectively.

In [20], Csiszár investigated the hypothesis testing problem between i.i.d. observations by considering the β-cutoff rate for the exponent of the best correct probability of type 1 with exponential constraint on the probability of type 2 error. More formally, he used the following definitions.

Definition 6.1 Fix $E > 0$. A rate r is called E-unachievable if there exists a sequence of acceptance regions $\mathcal{A}_n$ such that

$$\limsup_{n\to\infty} -\frac{1}{n}\log(1-\mu_n) \leq r \quad \text{and} \quad \liminf_{n\to\infty} -\frac{1}{n}\log\lambda_n \geq E.$$

Definition 6.2 The infimum of all E-unachievable rates is defined as:

$$D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) \triangleq \inf\{r > 0 : r \text{ is } E\text{-unachievable}\},$$

and $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) = 0$ if the above set is empty.

For $0 < r < D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}})$, every acceptable region $\mathcal{A}_n$ with $\liminf_{n\to\infty} -\frac{1}{n}\log\lambda_n \geq E$ satisfies $\mu_n > 1 - e^{-nr}$ for n infinitely often.

Definition 6.3 Fix $\beta > 0$. $R_0 \geq 0$ is a reverse β-achievable rate for the general hypothesis testing problem if

$$D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) \geq \beta(E - R_0)$$

for every $E > 0$. The reverse β-cutoff rate is defined as the infimum of all reverse β-achievable rates, and is denoted by $R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$.

However, in [29], Han investigated the general hypothesis testing problem between arbitrary sources with memory by considering the exponent of the best correct probability of type 2 with exponential constraint on the probability of type 1 error. More formally, he used the following definitions.

Definition 6.4 **[29]** Fix $r > 0$. A rate E is called r-unachievable if there exists a sequence of acceptance regions $\mathcal{A}_n$ such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mu_n \geq r \quad \text{and} \quad \limsup_{n\to\infty} -\frac{1}{n}\log(1-\lambda_n) \leq E.$$

Definition 6.5 **[29]** The infimum of all r-unachievable rates is denoted by $B_e^*(r|\mathbf{X}\|\bar{\mathbf{X}})$:

$$B_e^*(r|\mathbf{X}\|\bar{\mathbf{X}}) \overset{\triangle}{=} \inf\{E > 0 : E \text{ is } r\text{-unachievable}\},$$

and $B_e^*(r|\mathbf{X}\|\bar{\mathbf{X}}) = 0$ if the above set is empty.

Proposition 6.1 **[29]** Fix $r > 0$. For the general hypothesis testing problem, we have that

$$B_e^*(r|\mathbf{X}\|\bar{\mathbf{X}}) = \inf_{R\in\mathbb{R}}\{R + \bar{\rho}(R) + [r - \bar{\rho}(R)]^+\},$$

where

$$\bar{\rho}(R) \triangleq \lim_{n\to\infty} -\frac{1}{n}\log P_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \le R\right\},$$

and $[x]^+ = \max\{x, 0\}$, provided the limit defining $\bar{\rho}(R)$ exists, and for any $M > 0$, there exists $K > 0$ such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log P_{\bar{X}^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \ge K\right\} \ge M.$$

Remark 1: Note that Csiszár's and Han's definitions seem different at first glance. In our investigation, we realized that in order to establish our results on the reverse cutoff rate for general sources with memory, a formula for the reliability function of the type 1 probability of correct decoding, $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}})$, is needed. However, in [29], Han provided a formula for the reliability function of the type 2 probability of correct decoding, $B_e^*(r|\mathbf{X}\|\bar{\mathbf{X}})$. This turned out to be an obstacle, since we were not able to derive the reverse cutoff rate formula by directly using the formula for $B_e^*(r|\mathbf{X}\|\bar{\mathbf{X}})$. To overcome this obstacle, we observed that if we interchange the role of the null and alternative hypotheses distributions (i.e., $\mathbf{X} \leftrightarrow \bar{\mathbf{X}}$), and also r with E (i.e., $r \leftrightarrow E$) in Han's definitions (Definitions 6.4 and 6.5), then a formula for $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}})$ can be readily obtained from Han's result. More specifically, we have the following.

Definition 6.6 Fix $E > 0$. A rate r is called E-unachievable if there exists a sequence of acceptance regions $\mathcal{A}_n' = \mathcal{A}_n^c$ (complement of $\mathcal{A}_n$) such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log\lambda_n \ge E \quad \text{and} \quad \limsup_{n\to\infty} -\frac{1}{n}\log(1-\mu_n) \le r,$$

where

$$\lambda_n = \Pr\{\bar{X}^n \notin \mathcal{A}_n'\} = \Pr\{\bar{X}^n \in \mathcal{A}_n\} \quad \text{and} \quad \mu_n = \Pr\{X^n \in \mathcal{A}_n'\} = \Pr\{X^n \notin \mathcal{A}_n\}.$$

Definition 6.7 The infimum of all E-unachievable rates is given by

$$B_e^*(E|\bar{\mathbf{X}}\|\mathbf{X}) = \inf\{r > 0 : r \text{ is } E\text{-unachievable}\},$$

and $B_e^*(E|\bar{\mathbf{X}}\|\mathbf{X}) = 0$ if the above set is empty.

With Definitions 6.6 and 6.7, Proposition 6.1 becomes as follows.

Proposition 6.2 For any $E > 0$,

$$B_e^*(E|\bar{\mathbf{X}}\|\mathbf{X}) = \inf_{R\in\mathbb{R}} \left\{R + \rho(R) + [E - \rho(R)]^+\right\},$$

where

$$\rho(R) \triangleq \lim_{n\to\infty} -\frac{1}{n}\log P_{\bar{X}^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \le R\right\},$$

provided the limit defining $\rho(R)$ exists, and for any $M > 0$, there exists $K > 0$ such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log P_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \ge K\right\} \ge M.$$

Remark 2: We can now clearly observe that Definitions 6.6 and 6.1 are identical. This indicates that Han's $B_e^*(E|\bar{\mathbf{X}}\|\mathbf{X})$ is in fact Csiszár's $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}})$. Hence, using Definitions 6.1 and 6.2, Proposition 6.2 should be as follows.

Proposition 6.3 For any $E > 0$,

$$D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) = \inf_{R\in\mathbb{R}} \left\{R + \rho(R) + [E - \rho(R)]^+\right\},$$

where

$$\rho(R) = \lim_{n\to\infty} -\frac{1}{n}\log P_{\bar{X}^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \le R\right\},$$

provided the limit defining $\rho(R)$ exists, and for any $M > 0$, there exists $K > 0$ such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log P_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \ge K\right\} \ge M. \tag{6.1}$$

The above proposition is a key ingredient for our main results in the following section.

6.2 Hypothesis Testing Reverse β-Cutoff Rate

For clarity of presentation, we herein restate the definition of the reverse β-cutoff rate (which was already given in Definition 6.3).

Definition 6.8 Fix $\beta > 0$. $R_0 \ge 0$ is a reverse β-achievable rate for the general hypothesis testing problem if

$$D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) \ge \beta(E - R_0)$$

for every $E > 0$. The reverse β-cutoff rate is defined as the infimum of all reverse β-achievable rates, and is denoted by $R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$.

In the degenerate case where $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) = 0$, we have that $R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) = \infty$. We herein assume that $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}})$ is not identically 0 for all values of E and that the conditions of Proposition 6.3 are satisfied. A graphical illustration of $R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$ is given in Figure 6.1.

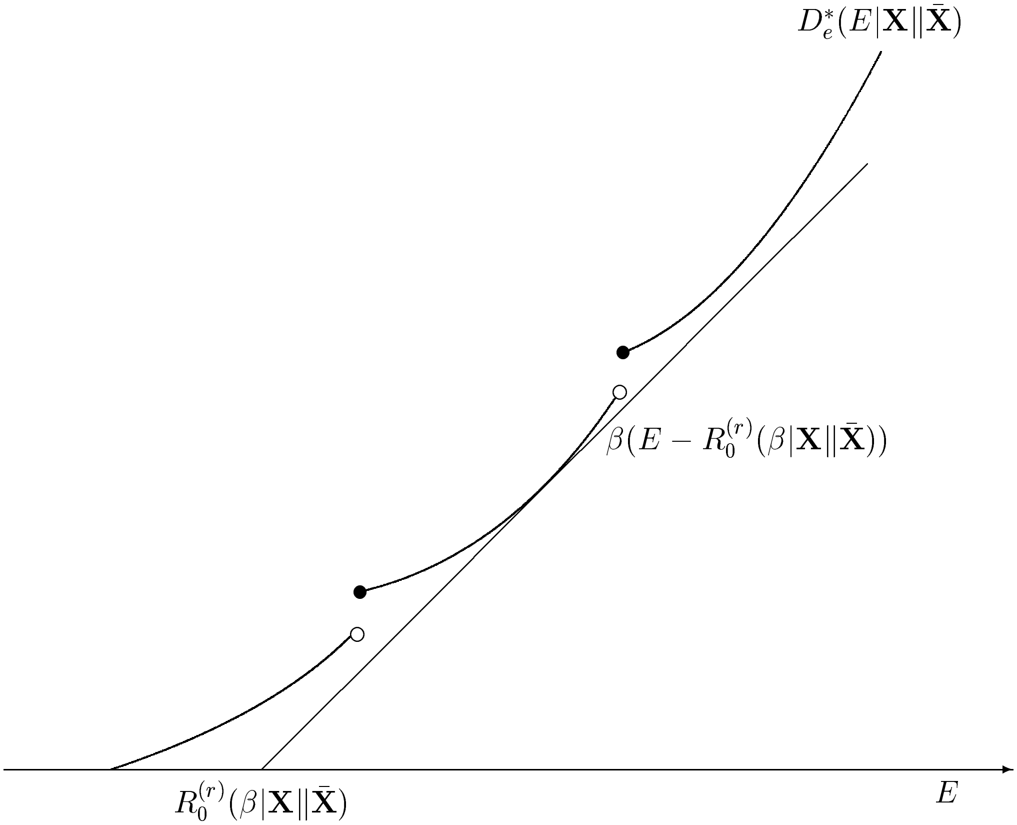
$D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}})$
$\beta(E - R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}))$
$R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$
E

Figure 6.1: A graphical illustration of the reverse β-cutoff rate, $R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$, for testing between two arbitrary sources $\mathbf{X}$ and $\bar{\mathbf{X}}$.

We first show the following lemmas, which will provide us the key mechanism to establish our reverse cutoff rate result.

Lemma 6.1 For all $E > 0$, we have that

$$D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) \leq E + \inf\{R \in \mathbb{R} : \rho(R) \leq E\}.$$

Proof: We have the following.

$$\begin{aligned} D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) &= \inf_{R\in\mathbb{R}} \left\{R + \rho(R) + [E - \rho(R)]^+\right\} \quad \text{(by Proposition 6.3)} \\ &= \min\left\{\inf_{\rho(R)\leq E}\{R + E\}, \inf_{\rho(R)>E}\{R + \rho(R)\}\right\} \\ &\leq \inf_{\rho(R)\leq E}\{R + E\} \\ &= E + \inf\{R \in \mathbb{R} : \rho(R) \leq E\}. \end{aligned}$$

□

Lemma 6.2 Assume that $\rho(R)$ is convex, and also assume that there exists an R such that $R + \rho(R) = 0$. Then for those E satisfying $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) > 0$,

$$D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) = E + \inf\{R \in \mathbb{R} : \rho(R) \leq E\}.$$

Proof: Since $\rho(R)$ is decreasing by definition and it is assumed to be convex, then it is continuous and strictly decreasing. Let R^* be the smallest one that satisfies $R + \rho(R) = 0$. Then for $E \leq \rho(R^*)$,

$$\begin{aligned} D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) &= \inf_{R\in\mathbb{R}} \left\{R + \rho(R) + [E - \rho(R)]^+\right\} \quad \text{(by Proposition 6.3)} \\ &\leq R^* + \rho(R^*) + [E - \rho(R^*)]^+ = 0. \end{aligned}$$

Hence, the set of values of E such that $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) > 0$ does not include $E \leq \rho(R^*)$. Now as $\rho(R)$ is assumed convex, its slope is strictly increasing, which implies that the slope of $\rho(R)$ is less than -1 for $R < R^*$. This immediately gives that the slope of the function $R + \rho(R)$ is negative for $R < R^*$. Consequently, for any $E > \rho(R^*)$ (which corresponds to $R < R^*$ since $\rho(R)$ is strictly decreasing),

$$\begin{aligned}
\inf_{\{R:\rho(R)>E\}} \{R + \rho(R)\} &= \{R + \rho(R)\}|_{R=\rho^{-1}(E)} \\
&= \rho^{-1}(E) + E = \inf_{\rho(R)\leq E} \{R + E\},
\end{aligned}$$

where

$$\rho^{-1}(E) \stackrel{\triangle}{=} \inf\{a : \rho(a) \leq E\},$$

is the quantile or inverse of $\rho(\cdot)$. Thus,

$$\begin{aligned}
D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) &= \inf_{R\in\mathbb{R}} \left\{R + \rho(R) + [E - \rho(R)]^+\right\} \\
&= \min\left\{ \inf_{\rho(R)\leq E}\{R + E\}, \inf_{\rho(R)>E}\{R + \rho(R)\} \right\} \\
&= \inf_{\rho(R)\leq E}\{R + E\} \\
&= E + \inf\{R \in \mathbb{R} : \rho(R) \leq E\}.
\end{aligned}$$

□

It is important to note that the above lemma does not necessarily hold in general; this is illustrated in the following example for the case where $\rho(R)$ is not convex.

Example 1: Let

$$\rho(R) = \begin{cases} 0, & R > 2; \\ -\frac{1}{2}R + 1, & -2 \leq R < 2; \\ -2R - 2, & -4 \leq R < -2; \\ -\frac{1}{2}R + 4, & -6 \leq R < -4; \\ -R + 1, & R < -6, \end{cases}$$

which is continuous and decreasing but not convex. Hence,

$$R + \rho(R) = \begin{cases} 0, & R > 2; \\ \frac{1}{2}R + 1, & -2 \leq R < 2; \\ -R - 2, & -4 \leq R < -2; \\ \frac{1}{2}R + 4, & -6 \leq R < -4; \\ 1, & R < -6, \end{cases}$$

Then indeed,

$$\begin{aligned} D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) &= \min\left\{ \inf_{\rho(R) \leq E} [R + E], \inf_{\rho(R) > E} [R + \rho(R)] \right\} \\ &= \inf_{\rho(R) > E} \{R + \rho(R)\} = \begin{cases} 0, & 0 < E \leq 2; \\ \frac{1}{2}E - 1, & 2 < E \leq 4; \\ 1, & E > 4, \end{cases} \end{aligned}$$

and

$$E + \inf\{R : \rho(R) \leq E\} = \begin{cases} -E + 2, & 0 < E \leq 2; \\ \frac{1}{2}E - 1, & 2 < E \leq 6; \\ -E + 8, & 6 < E \leq 7; \\ 1, & E > 7. \end{cases}$$

Lemma 6.3 Fix $t < 0$. Also, assume that $\rho(R)$ is convex, and suppose that there exists an R such that $R + \rho(R) = 0$. The following two conditions are equivalent.

$$(\forall\ R \in \mathbb{R}) \quad \rho(R) \geq -R(1-t) + tR_0 \tag{6.2}$$

and

$$(\forall\ E > 0) \quad D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) \geq \frac{t}{t-1}(E - R_0). \tag{6.3}$$

Proof:

a) (6.2)⇒(6.3). By Lemma 6.2, for those E satisfying $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) > 0$, we have that

$$\begin{aligned} D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) &= E + \inf\{R \in \mathbb{R} : \rho(R) \leq E\} \\ &\geq E + \inf\{R \in \mathbb{R} : -R(1-t) + tR_0 \leq E\} \\ &= \frac{t}{t-1}\left(E - R_0\right), \end{aligned}$$

where the inequality follows from (6.2). This implies that

$$\inf\{E > 0 : D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) > 0\} \leq R_0.$$

Hence, for these E satisfying $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) = 0$, the claim also holds since $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}})$ is increasing.

b) (6.3)⇒(6.2). By Lemma 6.1 and (6.3), for $E > 0$, we have that

$$\inf\{R \in \mathbb{R} : \rho(R) \leq E\} \geq \frac{t}{t-1}(E - R_0) - E = \frac{1}{t-1}E - \frac{t}{t-1}R_0.$$

Thus

$$E \leq \rho\left(\frac{1}{t-1}E - \frac{t}{t-1}R_0\right),$$

since $\rho(\cdot)$ is strictly decreasing. Letting

$$R = \frac{1}{t-1}E - \frac{t}{t-1}R_0,$$

or

$$E = -R(1-t) + tR_0,$$

the above inequality can be rewritten as

$$\rho(R) \geq -R(1-t) + tR_0,$$

where $R \in \mathbb{R}$. □

We next employ Lemma 6.3 to show our main result regarding the reverse cutoff rate.

Theorem 6.1 (Reverse β-cutoff rate formula). Assume that $\rho(R)$ is convex, and suppose that there exists an R such that $R + \rho(R) = 0$. For the general hypothesis testing problem,

$$R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) \le \limsup_{n\to\infty} \frac{1}{n} D_{1/(1-\beta)}(X^n\|\bar{X}^n) \quad \text{for } 0 < \beta < 1,$$

and

$$R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) \ge \limsup_{n\to\infty} \frac{1}{n} D_{1/(1-\beta)}(X^n\|\bar{X}^n) \quad \text{for } 0 < \beta < \beta_{\max},$$

where

$$\beta_{\max} = \sup\left\{\beta \in (0,1) : \limsup_{n\to\infty} \frac{1}{n} D_{1/(1-\gamma)}(X^n\|\bar{X}^n) < \infty \text{ for every } 0 < \gamma < \beta\right\},$$

and

$$D_\alpha(X^n\|\bar{X}^n) \triangleq \frac{1}{\alpha - 1} \log\left(\sum_{x^n\in\mathcal{X}^n} [P_{X^n}(x^n)]^\alpha [P_{\bar{X}^n}(x^n)]^{1-\alpha}\right)$$

is the n-dimensional Rényi α-divergence. Note that from the above two inequalities, $R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}})$ is indeed equal to the limsup $\frac{1}{1-\beta}$-divergence rate for $0 < \beta < \beta_{\max}$.

Proof:[1]

1. *Forward part:* $R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) \le \limsup_{n\to\infty} \frac{1}{n} D_{1/(1-\beta)}(X^n\|\bar{X}^n)$ for $0 < \beta < 1$.

   By the equivalence of conditions (6.2) and (6.3), it suffices to show that

$$(\forall R \in \mathbb{R})\ \rho(R) \ge -R(1-t) + t \cdot \limsup_{n\to\infty} \frac{1}{n} D_{1-t}(X^n\|\bar{X}^n).$$

[1] For the proof of the continuous alphabet case, the same remark given in Observation B (cf. Section 5.2) applies.

Consider the twisted distribution defined as:

$$\begin{aligned} P^{(t)}_{X^n}(x^n) &\triangleq \frac{[P_{\bar{X}}(x^n)]^t[P_{X^n}(x^n)]^{1-t}}{\sum_{\hat{x}^n\in\mathcal{X}^n}[P_{\bar{X}}(\hat{x}^n)]^t[P_{X^n}(\hat{x}^n)]^{1-t}} \\ &= \exp\left\{(t-1)\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} + tD_{1-t}(X^n\|\bar{X}^n)\right\}P_{\bar{X}^n}(x^n). \end{aligned} \tag{6.4}$$

Then for $t<0$,

$$\begin{aligned} & P_{\bar{X}^n}\left\{x^n\in\mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)}\le R\right\} \\ =\ & \sum_{\left\{x^n\in\mathcal{X}^n:\frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)}\le R\right\}} P_{\bar{X}^n}(x^n) \\ =\ & \sum_{\left\{x^n\in\mathcal{X}^n:\frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)}\le R\right\}} \exp\left\{(1-t)\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} - tD_{1-t}(X^n\|\bar{X}^n)\right\}P^{(t)}_{X^n}(x^n) \\ \le\ & \exp\left\{nR(1-t) - tD_{1-t}(X^n\|\bar{X}^n)\right\} \sum_{\left\{x^n\in\mathcal{X}^n:\frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)}\le R\right\}} P^{(t)}_{X^n}(x^n) \\ \le\ & \exp\left\{nR(1-t) - tD_{1-t}(X^n\|\bar{X}^n)\right\}. \end{aligned}$$

So,

$$\begin{aligned} \rho(R) &= \liminf_{n\to\infty} -\frac{1}{n}\log P_{\bar{X}^n}\left\{x^n\in\mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)}\le R\right\} \\ &\ge -R(1-t) + t\cdot\limsup_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) \\ &= -R(1-t) + t\cdot\limsup_{n\to\infty}\frac{1}{n}D_{1/(1-\beta)}(X^n\|\bar{X}^n), \text{ for } \beta\triangleq\frac{t}{t-1}\in(0,1). \end{aligned}$$

2. *Converse part:* $R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) \ge \limsup_{n\to\infty}\frac{1}{n}D_{1/(1-\beta)}(X^n\|\bar{X}^n)$ for $0<\beta<\beta_{\max}$.

By the equivalence of (6.2) and (6.3), it suffices to show the existence of $\bar{R}$ for any $\delta>0$ such that

$$\rho(\bar{R}) \le -\bar{R}(1-t) + t\left(\limsup_{n\to\infty}\frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) + \frac{(1-t)}{t}3\delta\right),$$

where $t = \beta/(\beta - 1) < 0$. Let $\mathcal{N}$ be a set of positive integers such that

$$\lim_{n\in\mathcal{N}, n\to\infty} \frac{1}{n} D_{1-t}(X^n \| \bar{X}^n) = \limsup_{n\to\infty} \frac{1}{n} D_{1-t}(X^n \| \bar{X}^n)$$

and define

$$\lambda \stackrel{\triangle}{=} \sup\{R \in \mathbb{R} : \rho^{(t)}(R) > 0\},$$

where

$$\rho^{(t)}(R) \stackrel{\triangle}{=} \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n} \log P^{(t)}_{X^n} \left\{ x^n \in \mathcal{X}^n : \frac{1}{n} \log \frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \le R \right\},$$

is the twisted large deviation spectrum of the normalized log-likelihood ratio with parameter t. It can be shown that λ satisfies $-\infty < \lambda \le 0$ (cf. Lemmas 6.4 and 6.5 in Section 6.3). We then note by definition of $\rho^{(t)}(\cdot)$ and the finiteness property of λ that for any $\delta > 0$, there exists $\epsilon > 0$ such that

$$\rho^{(t)}(\lambda - \delta) = \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n} \log P^{(t)}_{X^n} \left\{ x^n \in \mathcal{X}^n : \frac{1}{n} \log \frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \le \lambda - \delta \right\} > \epsilon > 0.$$

As a result,

$$P^{(t)}_{X^n} \left\{ x^n \in \mathcal{X}^n : \frac{1}{n} \log \frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} > \lambda - \delta \right\} \ge 1 - e^{-n\epsilon} \text{ for } n \in \mathcal{N} \text{ sufficiently large.}$$

On the other hand, define

$$\bar{\rho}^{(t)}(R) \stackrel{\triangle}{=} \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n} \log P^{(t)}_{X^n} \left\{ x^n \in \mathcal{X}^n : \frac{1}{n} \log \frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \ge R \right\}$$

and

$$\bar{\lambda} \stackrel{\triangle}{=} \inf\{R \in \mathbb{R} : \bar{\rho}^{(t)}(R) > 0\}.$$

Then by noting that

$$\log \frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} = -D_{1-t}(X^n \| \bar{X}^n) + \frac{1}{t} \log \frac{P^{(t)}_{X^n}(x^n)}{P_{X^n}(x^n)},$$

we have:

$$\bar{\rho}^{(t)}(R) = \sigma\left(tR + \frac{t}{n}D_{1-t}(X^n\|\bar{X}^n)\right),$$

where

$$\sigma(R) \stackrel{\triangle}{=} \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n}\log P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}^{(t)}(x^n)}{P_{X^n}(x^n)} \le R\right\},$$

and

$$\begin{aligned}\bar{\lambda} &= \frac{1}{t}\sup\{R\in\mathbb{R} : \sigma(R) > 0\} - \frac{1}{n}D_{1-t}(X^n\|\bar{X}^n) \\ &\le 0, \qquad (6.5)\end{aligned}$$

where (6.5) follows from Lemma 5.4 in Section 5.4, and the non-negativity [20] of the Rényi divergence $D_{1-t}(X^n\|\bar{X}^n)$. This indicates the existence of $\bar{\epsilon} > 0$ such that $\bar{\rho}^{(t)}(\delta) > \bar{\epsilon}$, which immediately gives that for $n \in \mathcal{N}$ sufficiently large,

$$P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \ge \delta\right\} \le e^{-n\bar{\epsilon}}.$$

Therefore, for $n \in \mathcal{N}$ sufficiently large,

$$\begin{aligned}&P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \delta > \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} > \lambda - \delta\right\} \\ \ge\ &P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} > \lambda - \delta\right\} \\ &-P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{X^n}(x^n)}{P_{\bar{X}^n}(x^n)} \ge \delta\right\} \\ \ge\ &1 - e^{-n\epsilon} - e^{-n\bar{\epsilon}}. \qquad (6.6)\end{aligned}$$

Let $I_1 \stackrel{\triangle}{=} (\lambda - \delta, b_1)$, and[2]

[2] Note that when $\lambda < 0$, $L \ge 2$; so the definition is well-established. However, in case $\lambda = 0$, we just take $L = 1$, and $I_1 = (-\delta, \delta)$.

$$I_k \triangleq [b_{k-1}, b_k) \text{ for } 2 \le k \le L \triangleq \left\lceil \frac{2\delta - \lambda}{2\delta} \right\rceil,$$

where $b_k \triangleq (\lambda - \delta) + 2k\delta$ for $1 \le k < L$, and $b_L \triangleq \delta$. By (6.6), there exists $1 \le k(n) \le L$ such that

$$P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \in I_{k(n)}\right\} \ge \frac{1 - e^{-n\epsilon} - e^{-n\bar{\epsilon}}}{L}, \tag{6.7}$$

for $n \in \mathcal{N}$ sufficiently large. Then, by letting $R_1 \triangleq \limsup_{n\in\mathcal{N}, n\to\infty} b_{k(n)} + \delta$, we obtain that for $n \in \mathcal{N}$ sufficiently large,

$$P_{\bar{X}^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \le R_1\right\} \ge P_{\bar{X}^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \in I_{k(n)}\right\}.$$

However, for sufficiently large $n \in \mathcal{N}$, we have that:

$$\begin{aligned}
& P_{\bar{X}^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \in I_{k(n)}\right\} \\
= \; & \sum_{\left\{x^n\in\mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)}\in I_{k(n)}\right\}} P_{\bar{X}^n}(x^n) \\
= \; & \sum_{\left\{x^n\in\mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)}\in I_{k(n)}\right\}} e^{-tD_{1-t}(X^n\|\bar{X}^n)} e^{(1-t)\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)}} P_{X^n}^{(t)}(x^n) \\
\ge \; & e^{-tD_{1-t}(X^n\|\bar{X}^n)} e^{(1-t)nb_{k(n)-1}} \sum_{\left\{x^n\in\mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)}\in I_{k(n)}\right\}} P_{X^n}^{(t)}(x^n) \\
= \; & e^{-tD_{1-t}(X^n\|\bar{X}^n)} e^{(1-t)nb_{k(n)-1}} P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \in I_{k(n)}\right\} \\
\ge \; & \frac{1 - e^{-n\epsilon} - e^{-n\bar{\epsilon}}}{L} e^{-tD_{1-t}(X^n\|\bar{X}^n)} e^{(1-t)nb_{k(n)-1}}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\rho(R_1) &= \liminf_{n\to\infty} -\frac{1}{n}\log P_{\bar{X}^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \le R_1\right\} \\
&\le \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n}\log P_{\bar{X}^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \le R_1\right\} \\
&\le t \limsup_{n\in\mathcal{N}, n\to\infty} \frac{1}{n} D_{1-t}(X^n\|\bar{X}^n) - (1-t)\limsup_{n\in\mathcal{N}, n\to\infty} b_{k(n)-1} \\
&\le t \limsup_{n\to\infty} \frac{1}{n} D_{1-t}(X^n\|\bar{X}^n) - (1-t)\limsup_{n\in\mathcal{N}, n\to\infty} b_{k(n)} + 2\delta(1-t) \\
&= t \limsup_{n\to\infty} \frac{1}{n} D_{1-t}(X^n\|\bar{X}^n) - (1-t)R_1 + 3\delta(1-t).
\end{aligned}$$

Since δ can be made arbitrarily small, the proof is completed. □

We observe that the conditions given in the above theorem are not necessary for the expression of the reverse β-cutoff rate to be given by the lim sup $\frac{1}{1-\beta}$-divergence rate. This is illustrated in the following example, where we show that $\rho(R)$ is not convex while

$$R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) = \limsup_{n\to\infty}\frac{1}{n}D_{1/(1-\beta)}(X^n\|\bar{X}^n) \quad \text{for } 0<\beta<1.$$

Example 2: Let $P_{\bar{X}^n}(a_n) = 1 - e^{-2n}$ and $P_{\bar{X}^n}(b_n) = e^{-2n}$, where $a_n \ne b_n$ and $a_n, b_n \in \mathcal{X}^n$. Also, let $P_{X^n}(a_n) = 1 - e^{-cn}$ and $P_{X^n}(b_n) = e^{-cn}$, where $0 < c < 2$. Then, the log-likelihood ratio, Z_n, is given by

$$Z_n = \log\frac{P_{\bar{X}^n}(X^n)}{P_{X^n}(X^n)} = \begin{cases} \log\dfrac{1-e^{-2n}}{1-e^{-cn}}, & \text{with probability (in } P_{\bar{X}^n}\text{) } 1-e^{-2n} \\ \\ -(2-c)n, & \text{with probability (in } P_{\bar{X}^n}\text{) } e^{-2n}, \end{cases}$$

which implies that

$$\rho(R) \quad = \quad \lim_{n\to\infty} -\frac{1}{n}\log \Pr\left\{\frac{1}{n}Z_n \le R\right\} = \begin{cases} 0, & \text{for } R \ge 0 \\ 2, & \text{for } -(2-c) \le R < 0 \\ \infty, & \text{for } R < -(2-c). \end{cases}$$

Note that $\rho(R)$ in not convex but $R + \rho(R) = 0$ for $R = 0$. Note also that Han's condition (6.1) is satisfied since $P_{X^n}(\cdot)$ and $P_{\bar{X}^n}(\cdot)$ are absolutely continuous with respect to each other. Let us first compute the α-divergence rate between X^n and $\bar{X}^n$, where $\alpha > 1$. The normalized n-dimensional α-divergence is given by

$$\frac{1}{n}D_\alpha(X^n\|\bar{X}^n) = \frac{1}{n(\alpha-1)}\log\left[(1-e^{-cn})^\alpha(1-e^{-2n})^{1-\alpha} + e^{-cn\alpha}e^{-2n(1-\alpha)}\right].$$

We have the following three cases.

1. $c\alpha + 2 - 2\alpha > 0$. Note that e^{-cn} and e^{-2n} approach 0 as $n \to \infty$ and that $e^{-cn\alpha}e^{-2n(1-\alpha)} = e^{-n(c\alpha+2-2\alpha)}$, which also approaches 0 as $n \to \infty$. Hence, the α-divergence rate is equal to 0 since the argument of the logarithm $\to$ 1 as $n \to \infty$.

2. $c\alpha + 2 - 2\alpha < 0$. In this case, since $e^{-n(c\alpha+2-2\alpha)} \to \infty$ as $n \to \infty$, the argument of the logarithm, for large n, is dominated by $e^{-n(c\alpha+2-2\alpha)}$. Hence

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{n}D_\alpha(X^n\|\bar{X}^n) &= \lim_{n\to\infty} -\frac{n(c\alpha+2-2\alpha)}{n(\alpha-1)} \\ &= \frac{c\alpha+2-2\alpha}{1-\alpha}. \end{aligned}$$

3. $c\alpha + 2 - 2\alpha = 0$. Clearly, the α-divergence rate is equal to 0 in this case.

Let us now compute the reverse β-cutoff rate. First, we need to compute $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}})$ using Proposition 6.3. We have the following cases.

- $E > 2$. We have that

$$R+\rho(R)+[E-\rho(R)]^+ \;=\; \begin{cases} R+E, & \text{for } R \geq 0 \\ R+E, & \text{for } -(2-c) \leq R < 0 \\ \infty, & \text{for } R < -(2-c). \end{cases}$$

  Hence

$$\begin{aligned} D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) &= \inf_{R\in\mathbb{R}} \left\{R+\rho(R)+[E-\rho(R)]^+\right\} \\ &= E-2+c. \end{aligned}$$

- $0 < c < E \leq 2$. In this case

$$R+\rho(R)+[E-\rho(R)]^+ \;=\; \begin{cases} R+E, & \text{for } R \geq 0 \\ R+2, & \text{for } -(2-c) \leq R < 0 \\ \infty, & \text{for } R < -(2-c). \end{cases}$$

  Hence, $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) = c$.

- $0 < E \leq c$. In this case

$$R+\rho(R)+[E-\rho(R)]^+ \;=\; \begin{cases} R+E, & \text{for } R \geq 0 \\ R+2, & \text{for } -(2-c) \leq R < 0 \\ \infty, & \text{for } R < -(2-c). \end{cases}$$

  Hence, $D_e^*(E|\mathbf{X}\|\bar{\mathbf{X}}) = E$.

The reverse β-cutoff rate is the E-axis intercept of the line of slope β passing by the point $(2, c)$ as illustrated in Figure 6.2. By straightforward calculation, we get that

$$R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) = -\frac{c}{\beta} + 2.$$

For $\alpha = 1/(1-\beta)$, we get that

$$R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) = \frac{c\alpha + 2 - 2\alpha}{1-\alpha}\cdot$$

Since, by definition, $R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) \geq 0$, it is straightforward to check that

$$R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) = \limsup_{n\to\infty} \frac{1}{n} D_{1/(1-\beta)}(X^n\|\bar{X}^n) \quad \text{for } 0 < \beta < 1.$$

Note that for this example, since the α-divergence rate is always finite , it follows directly that $\beta_{\max} = 1$.

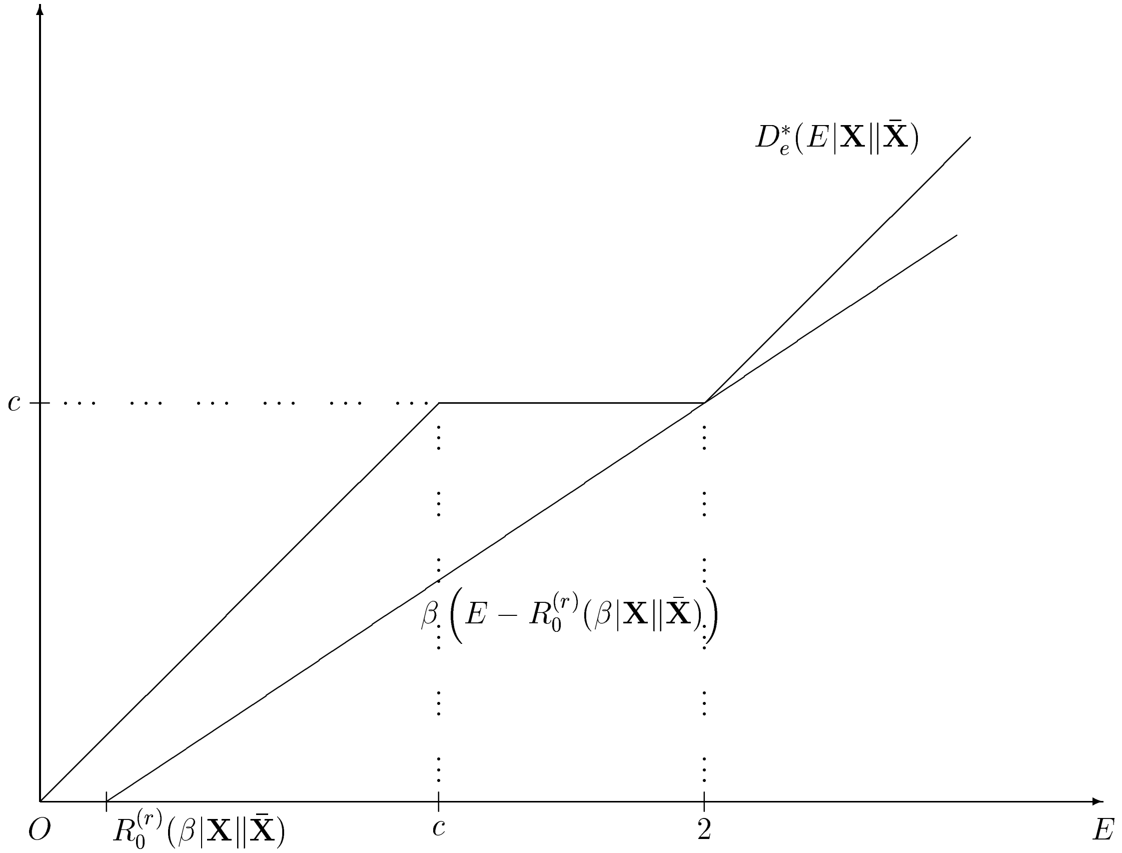

Figure 6.2: Reliability function of the type 1 probability of correct decoding for testing between the two sources $P_{X^n}(\cdot)$ and $P_{\bar{X}^n}(\cdot)$ as given in Example 1.

We next show that in the case of i.i.d. finite-alphabet observations, our result in Theorem 6.1 reduces to Csiszár's result [20]; i.e., the reverse β-cutoff rate is given by the Rényi divergence with parameter $\frac{1}{1-\beta}$, for $0 < \beta < 1$.

Corollary 6.1 For the hypothesis testing problem between two finite-alphabet memoryless sources $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ and $\bar{\mathbf{X}} = \{\bar{X}_i\}_{i=1}^{\infty}$, we have that

$$R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) = D_{1/(1-\beta)}(X\|\bar{X}) \quad \text{for } 0 < \beta < 1.$$

Proof: By Cramer's theorem [12, p. 9], we get that

$$\begin{aligned}\rho(R) &= \inf_{s\in(-\infty,R]} I_Z(s) \\ &= \begin{cases} I_Z(R), & R < E_{P_{\bar{X}}}[Z] \\ 0, & \text{otherwise,}\end{cases}\end{aligned}$$

where $E_{P_{\bar{X}}}[Z]$ denotes the expectation of the log-likelihood ratio $Z = \log\frac{P_{\bar{X}}(\bar{X})}{P_X(\bar{X})}$ with respect to $P_{\bar{X}}$, and

$$I_Z(s) = \sup_{\theta\in\mathbb{R}} \left(\theta s - \log M_Z(\theta)\right),$$

where $M_Z(\theta) = E_{P_{\bar{X}}}[exp\{\theta Z\}]$ is the moment generating function of the random variable Z. Clearly, $\rho(R)$ is convex [12, p. 9], and it is infinite[3] when $R < \log m$, where

$$m \triangleq \min\left\{\frac{P_{\bar{X}}(x)}{P_X(x)}, x \in \mathcal{X}\right\}.$$

[3] Indeed, let $R = \log m - \delta$, for some positive constant δ. Then

$$\theta R - \log M_Z(\theta) = -\theta\delta + \log\frac{m^\theta}{\sum_x P_{\bar{X}}(x)\left[\frac{P_{\bar{X}}(x)}{P_X(x)}\right]^\theta},$$

which diverges to $+\infty$ when $\theta \to -\infty$, since the last term converges to a constant.

Let us now prove that there exists an R such that $\rho(R) + R = 0$. If we differentiate $(\theta R - \log M_Z(\theta))$ with respect to θ, and set the result to 0, we get that

$$R = \frac{M'_Z(\theta)}{M_Z(\theta)} \stackrel{\triangle}{=} f(\theta). \tag{6.8}$$

By Schwarz inequality, it is straightforward to verify that the function $f(\theta)$ is strictly increasing[4]. Hence, f^{-1} exists and is differentiable ($f'(\theta) > 0$, for all $\theta \in \mathbb{R}$). Note that

$$f(\theta) \in I \stackrel{\triangle}{=} [\log m, \log M],$$

where

$$M \stackrel{\triangle}{=} \max \left\{ \frac{P_{\bar{X}}(x)}{P_X(x)}, x \in \mathcal{X} \right\}.$$

Note also that $E_{P_{\bar{X}}}[Z] \leq \log M$. Therefore, for every $R \in [\log m, E_{P_{\bar{X}}}[Z]]$, there exists a unique θ which satisfies equation (6.8). Hence,

$$\rho(R) = f^{-1}(R)R - \log M_Z(f^{-1}(R)),$$

which yields that $\rho(R)$ is differentiable. Since $\rho(R)$ is infinite when $R < \log m$ and is equal to 0 for $R \geq E[Z]$, the set of slopes of tangent lines to $\rho(R)$ is between $-\infty$

[4] We have that

$$f(\theta) = \frac{\sum_x P_{\bar{X}}(x) \left[\frac{P_{\bar{X}}(x)}{P_X(x)}\right]^\theta \log \frac{P_{\bar{X}}(x)}{P_X(x)}}{\sum_x P_{\bar{X}}(x) \left[\frac{P_{\bar{X}}(x)}{P_X(x)}\right]^\theta},$$

and hence

$$f'(\theta) = \frac{\sum_x P_{\bar{X}}(x) \left[\frac{P_{\bar{X}}(x)}{P_X(x)}\right]^\theta \left(\log \frac{P_{\bar{X}}(x)}{P_X(x)}\right)^2 \cdot \sum_x P_{\bar{X}}(x) \left[\frac{P_{\bar{X}}(x)}{P_X(x)}\right]^\theta - \left(\sum_x P_{\bar{X}}(x) \left[\frac{P_{\bar{X}}(x)}{P_X(x)}\right]^\theta \log \frac{P_{\bar{X}}(x)}{P_X(x)}\right)^2}{\left(\sum_x P_{\bar{X}}(x) \left[\frac{P_{\bar{X}}(x)}{P_X(x)}\right]^\theta\right)^2}.$$

By Schwarz inequality, $f'(\theta) \geq 0$ with equality iff $P_{\bar{X}}(x) = cP_X(x)$ for all $x \in \mathcal{X}$ where c is some positive constant. Thus, $f'(\theta) > 0$, since in the hypothesis testing problem it is assumed that the sources are different.

and 0. Hence, there exists a tangent line with slope -1 to $\rho(R)$. Let R^* be the point where the line of slope -1 is tangent to $\rho(R)$. By definition

$$\rho(R^*) = \sup_{\theta \in \mathbb{R}} \left(\theta R^* - \log M_Z(\theta) \right).$$

If this supremum is achieved by some $\theta^* \neq -1$, it would contradict the fact that $\theta R^* - \log M_Z(\theta)$ is a lower bound for $\rho(R^*)$ for each θ (any line with slope different of -1 passing through the point $(R^*, \rho(R^*))$ cannot be a lower bound to $\rho(R^*)$). Hence

$$\rho(R^*) = -R^* - \log M_Z(\theta)\big|_{\theta=-1}.$$

But $M_Z(\theta) = 1$ for $\theta = -1$, hence $\rho(R^*) = -R^*$. Hence, there exists an R such that $R + \rho(R) = 0$. Finally, we show that $\beta_{\max} = 1$. Note first that

$$\limsup_{n\to\infty} \frac{1}{n} D_\alpha(X^n \| \bar{X}^n) = D_\alpha(X \| \bar{X}).$$

If $\beta_{\max} < 1$, then there exists some $\alpha > 1$ such that $D_\alpha(X\|\bar{X}) = \infty$. Since the alphabet $\mathcal{X}$ is finite, this implies that $\sum_x p_X^\alpha(x) p_{\bar{X}}^{1-\alpha}(x)$ is infinite. Hence, there exists at least an $x \in \mathcal{X}$ such that $P_X(x) \neq 0$ and $P_{\bar{X}}(x) = 0$. But this certainly violates Han's condition (6.1) in Theorem 6.3. Hence $\beta_{\max} = 1$ and the corollary is proved.

□

We finally present a class of sources with memory for which the reverse β-cutoff rate is given by the Rényi $\frac{1}{1-\beta}$-divergence rate for all $0 < \beta < 1$.

Corollary 6.2 Consider the hypothesis testing problem between finite-alphabet sources with memory such that the log-likelihood ratio process $\{Z_n\}$, where $Z_n = \log \frac{P_{\bar{X}^n}(\bar{X}^n)}{P_{X^n}(\bar{X}^n)}$, satisfies both hypotheses of the Gärtner-Ellis Theorem [12, p. 15]:

- $\phi(\theta) \stackrel{\triangle}{=} \lim_{n\to\infty} \frac{1}{n}\phi_n(\theta)$ exists for all $\theta \in \mathbb{R}$,
- ϕ is differentiable on d_φ, where $d_\varphi \stackrel{\triangle}{=} \{\theta : \phi(\theta) < \infty\}$,

where $\phi_n(\theta) \stackrel{\triangle}{=} \log E_{P_{\bar{X}^n}}[\exp(\theta Z_n)]$. Also, suppose that $\frac{1}{n}\phi_n(\theta)$ converges uniformly in n to $\phi(\theta)$. Then the reverse β-cutoff rate satisfies

$$R_0^{(r)}(\beta|\mathbf{X}\|\bar{\mathbf{X}}) = \lim_{n\to\infty} \frac{1}{n} D_{1/(1-\beta)}(X^n\|\bar{X}^n) \quad \text{for } 0 < \beta < 1.$$

Proof: To prove the result, we need to show that for sources satisfying the Gärtner-Ellis Theorem, the Rényi divergence rate exists, that the conditions of Theorem 6.1 hold and that $\beta_{\max} = 1$. First, the Rényi divergence rate exists from the first hypothesis of the Gärtner-Ellis Theorem and the fact that

$$\frac{1}{n} D_{\frac{1}{1-\beta}}(X^n\|\bar{X}^n) = \frac{1-\beta}{\beta}\frac{1}{n}\phi_n\left(\frac{1}{\beta - 1}\right).$$

Next, by the Gärtner-Ellis Theorem, we have that

$$\rho(R) = \sup_{\theta\in\mathbb{R}} \{\theta R - \phi(\theta)\}.$$

Clearly, $\rho(R)$ is convex in R. Let us show that there exists an R such that $R + \rho(R) = 0$. In order to employ the previous corollary, we let

$$\rho_n(R) \stackrel{\triangle}{=} \sup_{\theta\in\mathbb{R}} \left\{\theta R - \frac{1}{n}\phi_n(\theta)\right\},$$

for $n = 1, 2, \ldots$. Along the same lines as in the previous corollary, it can be shown that there exists an R_n^* such that $R_n^* + \rho_n(R_n^*) = 0$, $n = 1, 2, \ldots$ On the other hand, $|\phi(\theta) - \frac{1}{n}\phi_n(\theta)| < \delta_n$ for n sufficiently large, where $\delta_n > 0$ is independent of θ by the uniform convergence assumption, and converges to 0 as $n \to \infty$ for all $\theta \in \mathbb{R}$. Therefore

$$|\rho_n(R) - \rho(R)| < \sup_{\theta \in \mathbb{R}} \delta_n = \delta_n, \tag{6.9}$$

for all $R \in \mathbb{R}$. In particular, (6.9) holds for $R = R_n^*$:

$$|\rho_n(R_n^*) - \rho(R_n^*)| < \delta_n.$$

But $\rho_n(R_n^*) + R_n^* = 0$, therefore $|\rho(R_n^*) + R_n^*| < \delta_n$. Define

$$R^* \triangleq \limsup_{n \to \infty} R_n^*.$$

We conclude that $\rho(R^*) + R^* = 0$. Finally, the fact that $\beta_{\max} = 1$ follows directly from the first hypothesis of the Gärtner-Ellis Theorem.

□

Numerical Examples: We briefly present two examples of memoryless sources where we explicitly verify the existence of R such that $R + \rho(R) = 0$.

Example 3: *Finite-alphabet memoryless sources:* Consider Example 1 in Section 5.3 where $\mathbf{X}$ and $\bar{\mathbf{X}}$ are interchanged. Note that $\rho(R)$ is equal to $\eta(R)$ in this case. It is straightforward to check that $R + \rho(R) = 0$ for R approximately -0.13.

Example 4: *Continuous alphabet memoryless sources:* Consider Example 2 in Section 5.3 where $\mathbf{X}$ and $\bar{\mathbf{X}}$ are interchanged. Note that $\rho(R)$ is equal to $\eta(R)$ in this case. By straightforward calculation we get that $R + \rho(R) = 0$ for $R = -2\nu^2$.

6.3 Properties of $\boldsymbol{\lambda}$

Lemma 6.4 For $t < 0$, $\lambda \leq 0$.

Proof: Observe that for $R > 0$,

$$\begin{aligned}
& P_{X^n}^{(t)}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} > R\right\} \\
\leq\ & e^{-nR(1-t)+tD_{1-t}(X^n\|\bar{X}^n)} P_{\bar{X}^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} > R\right\} \\
\leq\ & e^{-nR(1-t)+tD_{1-t}(X^n\|\bar{X}^n)} \\
\leq\ & e^{-nR(1-t)},
\end{aligned}$$

where the last inequality follows from the non-negativity of $D_{1-t}(X^n\|\bar{X}^n)$. This implies that for $R > 0$,

$$\rho^{(t)}(R) \ \leq \ \liminf_{n\in\mathcal{N}, n\to\infty} -\frac{1}{n}\log\left(1 - e^{-nR(1-t)}\right) = 0,$$

which immediately implies that $\lambda \leq 0$. □

Lemma 6.5 For $0 > t > \beta_{\max}/(\beta_{\max} - 1)$, $\lambda > -\infty$.

Proof: If $\lambda = -\infty$, then $\rho^{(t)}(R) = 0$ for every $R \in \mathbb{R}$. Hence, by choosing any $\delta > 0$ satisfying $t > t - \delta > \beta_{\max}/(\beta_{\max} - 1)$, we have:

$$\begin{aligned} & P^{(t)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \le R\right\} \\ \le\;\; & e^{tD_{1-t}(X^n\|\bar{X}^n)-(t-\delta)D_{1-(t-\delta)}(X^n\|\bar{X}^n)+\delta nR}\, P^{(t-\delta)}_{X^n}\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\log\frac{P_{\bar{X}^n}(x^n)}{P_{X^n}(x^n)} \le R\right\} \\ \le\;\; & e^{-(t-\delta)D_{1-(t-\delta)}(X^n\|\bar{X}^n)+\delta nR}, \end{aligned}$$

which implies that

$$0 = \rho^{(t)}(R) \ge (t-\delta) \limsup_{n\in\mathcal{N}, n\to\infty} \frac{1}{n} D_{1-(t-\delta)}(X^n\|\bar{X}^n) - \delta R.$$

This indicates that

$$\limsup_{n\to\infty}\frac{1}{n}D_{1-(t-\delta)}(X^n\|\bar{X}^n) \ge \limsup_{n\in\mathcal{N}, n\to\infty}\frac{1}{n}D_{1-(t-\delta)}(X^n\|\bar{X}^n) \ge \frac{\delta}{t-\delta}R \quad \text{for every } R \in \mathbb{R},$$

or equivalently,

$$\limsup_{n\to\infty}\frac{1}{n}D_{1-(t-\delta)}(X^n\|\bar{X}^n) = \infty,$$

which contradicts the assumption on $\beta_{\max}$. □

Chapter 7

Conclusion and Future Work

7.1 Summary

This thesis consists of two major parts.

In the first part, we studied Shannon's and Rényi's measure rates for finite-alphabet time-invariant Markov sources of arbitrary order and arbitrary initial distributions. We obtained computable expressions for the Kullback-Leibler divergence rate and the α-divergence rate between Markov sources. We also showed that their rate of convergence is of the order $1/n$. We also provided sufficient conditions under which the α-divergence rate reduces to the Kullback-Leibler divergence rate as $n \to \infty$ and $\alpha \to 1$. We obtained similar results for the Shannon entropy rate and the Rényi entropy rate. The main tools used in obtaining these results are the theory of non-negative matrices and Perron-Frobenius theory. As an application to hypothesis testing, we provided a simple proof of Stein's Lemma for irreducible stationary

Markov sources which goes along the same lines as in the i.i.d. case. As an application to source coding, we generalized Campbell's variable-length source coding theorem for i.i.d. sources to Markov sources.

In the second part, we examined the forward and reverse β-cutoff rates for the hypothesis testing problem between arbitrary sources with memory (not necessarily Markovian, ergodic, stationary, etc.) of arbitrary alphabet (countable or uncountable). We showed that the forward β-cutoff rate is given by the $\liminf$ α-divergence rate, where $\alpha = \frac{1}{1-\beta}$ and $\beta < 0$. Under two conditions on the log likelihood ratio large deviation spectrum, $\rho(R)$, we showed that the reverse β-cutoff rate is given by the $\limsup$ α-divergence rate, where $\alpha = \frac{1}{1-\beta}$ and $0 < \beta < \beta_{\max}$. For $\beta_{\max} \leq \beta < 1$, we provided an upper bound on the reverse cutoff rate. In particular, we examined i.i.d. observations and sources that satisfy the hypotheses of the Gärtner-Ellis Theorem. We showed that in these cases, the conditions on $\rho(R)$ are satisfied and that the reverse cutoff rate admits a simple form. We also provided several numerical examples to illustrate our forward and reverse cutoff rate results. The main tools used in obtaining these results are large deviation theory and the information spectrum approach.

7.2 Future Work

One possible direction for future work is the investigation of Shannon's and Rényi's information measure rates for general sources with memory (not necessarily Markovian), including hidden Markov sources. For instance, to the best of our knowledge, it is not known whether the Rényi entropy rate for finite-alphabet stationary ergodic sources exists or not. Further investigation of the reverse cutoff rate is also of interest. One aim is to investigate if the reverse β-cutoff rate result of Theorem 6.1 holds without any restriction on $\rho(R)$. Another direction is to study Csiszár's channel coding cutoff rates [20] for arbitrary discrete channels with memory using our information spectrum techniques.

Bibliography

[1] J. Aczél and Z. Daróczy, *On Measures of Information and Their Characterization*, Academic Press, New York, 1975.

[2] F. Alajaji and T. Fuja, "A communication channel modeled on contagion," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 2035–2041, November 1994.

[3] F. Alajaji, P.-N. Chen and Z. Rached, "Csiszár's cutoff rates for the general hypothesis testing problem," under preparation.

[4] F. Alajaji, P.-N. Chen and Z. Rached, "Csiszár's forward cutoff rate for testing between two arbitrary sources," *Proc. IEEE International Symposium on Information Theory*, Lausanne, Switzerland, June 29–July 5, 2002.

[5] V. Anantharam, "A large deviations approach to error exponents in source coding and hypothesis testing," *IEEE Transactions on Information Theory*, vol. 36, no. 4, pp. 938–943, July 1990.

[6] E. Arikan, “An inequality on guessing and its application to sequential decoding,” *IEEE Transactions on Information Theory*, vol. 42, No. 1, pp. 99–105, January 1996.

[7] R. B. Ash, *Information Theory*, Dover Publications, New York, 1965.

[8] M. B. Bassat and J. Raviv, “Rényi’s entropy and the probability of error,” *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 324–330, May 1978.

[9] M. B. Bassat, “f-entropies, probability of error, and feature selection,” *Information and Control*, vol. 39, pp. 227–242, 1978.

[10] P. Billingsley, *Ergodic Theory and Information*, John Wiley & Sons, Inc., New York, 1965.

[11] A. C. Blumer and R. J. McEliece, “The Rényi redundancy of generalized Huffman codes,” *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 1242–1249, September 1988.

[12] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*, Wiley, New York, 1990.

[13] L. L. Campbell, “A coding theorem and Rényi’s entropy,” *Information and Control*, vol. 8, pp. 423–429, 1965.

[14] P.-N. Chen and F. Alajaji, “Csiszár’s cutoff rates for arbitrary discrete sources,” *IEEE Transactions on Information Theory*, vol. 47, pp. 330–338, January 2001.

[15] P.-N. Chen, “General formulas for the Neyman-Pearson type 2 error exponent subject to fixed and exponential type 1 error bounds,” *IEEE Transactions on Information Theory*, vol. 42, January 1996.

[16] C. H. Chen, *Statistical Pattern Recognition*, Rochelle Park, NJ: Hayden Book Co., Ch. 4, 1973.

[17] C. K. Chow and C. N. Liu, “Approximating discrete probability distributions with dependence trees,” *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, May 1968.

[18] T. M. Cover and J. Thomas, *Elements of Information Theory*, John Wiley and Sons, Inc., 1991.

[19] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*, Methuen and Co Ltd, 1965.

[20] I. Csiszár, “Generalized cutoff rates and Rényi's information measures,” *IEEE Transactions on Information Theory*, vol. 41, pp. 26–34, January 1995.

[21] I. Csiszár and G. Longo, “On the error exponent for source coding and for testing simple statistical hypotheses,” *Studia Scientiarum Mathematicarum Hungarica*, vol. 6, pp. 181–191, 1971.

[22] Minh N. Do, “Fast approximation of Kullback-Leibler Distance for dependence trees and hidden Markov models,” submitted to *IEEE Signal Processing Letters.*

[23] U. Erez and R. Zamir, “Error exponents of modulo-additive noise channels with side information at the transmitter,”*IEEE Transactions on Information Theory*, vol. 47, pp. 210–218, January 2001.

[24] R. G. Gallager, *Discrete Stochastic Processes*, Kluwer, Boston, 1996.

[25] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, 1968.

[26] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, New York, 1990.

[27] T. S. Han and S. Verdú, “Approximation theory of output statistics,” *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 752–772, May 1993.

[28] T. S. Han, “The reliability functions of the general source with fixed-length coding,” *IEEE Transactions on Information Theory*, vol. 46, no. 6, pp. 2117–2132, September 2000.

[29] T. S. Han, “Hypothesis testing with the general source,” *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2415–2427, November 2000.

[30] T. S. Han, *Information-Spectrum Methods in Information Theory*, Tokyo, Japan: BaifuKan, 1998 (in Japanese).

[31] W. Hoeffding, “Asymptotically optimal test for multinomial distributions,” *Annals of Mathematical Statistics*, vol. 36, pp. 369–400, 1965.

[32] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[33] T. C. Hu, D. J. Kleitman and J. K. Tamaki, "Binary trees optimum under various criteria," *Siam Journal on Applied Mathematics*, vol. 37, no. 2, pp. 246–256, October 1979.

[34] F. Jelinek, "Buffer overflow in variable length coding of fixed rate sources," *IEEE Transactions on Information Theory,* vol. 14, pp. 490–501, May 1968.

[35] F. Jelinek, *Probabilistic Information Theory*, McGraw-Hill, New York, 1968.

[36] T. T. Kadota and L. A. Shepp, "On the best finite set of linear observables for discriminating two Gaussian signals," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 278–284, Apr. 1967.

[37] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, Feb. 1967.

[38] D. Kazakos and T. Cotsidas, "A decision theory approach to the approximation of discrete probability densities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 1, pp. 61–67, Jan. 1980.

[39] L. H. Koopmans, "Asymptotic rate of discrimination for Markov processes," *Annals of Mathematical Statistics*, vol. 31, pp. 982–994, 1960.

[40] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[41] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd edition, Academic, Toronto, 1985.

[42] K. Marton and P. C. Shields, "The positive-divergence and blowing-up properties," *Israel Journal of Mathematics*, vol. 86, 331–348, 1994.

[43] S. Natarajan, "Large deviations, hypotheses testing, and source coding for finite Markov chains," *IEEE Transactions on Information Theory*, vol. 31, pp. 360–365, May 1985.

[44] T. Nemetz, "On the α-divergence rate for Markov-dependent hypotheses," *Problems of Control and Information Theory,* vol. 3 (2), pp. 147–155, 1974.

[45] T. Nemetz, *Information Type Measures and Their Applications to Finite Decision-Problems*, Carleton Mathematical Lecture Notes, no. 17, May 1977.

[46] L. Pronzato, H. P. Wynn and A. A. Zhigljavsky, "Using Rényi entropies to measure uncertainty in search problems," *Lectures in Applied Mathematics*, vol. 33, pp. 253–268, 1997.

[47] Z. Rached F. Alajaji and L. L. Campbell, "Rényi's entropy rate for discrete Markov sources," *Proc. Conference on Information Sciences and Systems*, Baltimore, MD, March 17–19, 1999.

[48] Z. Rached, F. Alajaji, and L. L. Campbell, "Rényi's divergence and entropy rates for finite-alphabet Markov sources," *IEEE Transactions on Information Theory*, vol. 47, pp. 1553–1561, May 2001.

[49] Z. Rached, F. Alajaji and L. L. Campbell, “On the Kullback-Leibler divergence rate for finite-alphabet Markov sources,” *IEEE Transactions on Information Theory*, submitted.

[50] Z. Rached, F. Alajaji and L. L. Campbell, “A formula of the divergence rate for a class of Markov sources,” *Proc. Conference on Information Sciences and Systems*, Princeton, March 20–22, 2002.

[51] Z. Rached, F. Alajaji and L. L. Campbell, “On the Rényi divergence rate for finite-alphabet Markov sources,” *Proc. International Symposium on Information theory and Its Applications*, Honolulu, Hawaii, November 5–8, 2000.

[52] A. Rényi, “On measures of entropy and information,” *Proc. Fourth Berkeley Symposium on Mathematics and Statistics* **1**, 547–561, University of California Press, Berkeley, 1961.

[53] V. I. Romanovsky, *Discrete Markov Chains*, Wolters-Noordhoff Publishing, Groningen, 1970.

[54] E. Seneta, *Non-Negative Matrices and Markov Chains*, Springer-Verlag New York Inc., 1981.

[55] C. E. Shannon, “A mathematical theory of communications,” *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[56] P. C. Shields, “Two divergence-rate counterexamples,” *Journal of Theoretical Probability*, vol. 6, 521–545, 1993.

[57] I. J. Taneja and R. M. Capocelli, “On some inequalities and generalized entropies: A unified approach,” *Cybernetics and Systems*, vol. 16, 341–376, 1985.

[58] S. Verdú and T. S. Han, “A general formula for channel capacity,” *IEEE Transactions on Information Theory*, vol. 40, no. 4, July 1994.

[59] Z. Ye and T. Berger, *Information Measures For Discrete Random Fields*, Science Press, Beijing, New York, 1998.

VITA

ZIAD RACHED

EDUCATION:

- **Ph.D. in Mathematics (Communications)**, September 1998–August 2002. Queen's University, Canada.

- **M.Sc. in Applied Mathematics (Communications)**, September 1997–September 1998, Queen's University, Canada.

- **M.Sc. in Mathematics (Algebra)**, September 1994–May 1996, University of Ottawa, Canada.

- **B.Sc. in Mathematics and Sciences (Physics)**, September 1990–May 1994, University of Ottawa, Canada.

AWARDS AND HONORS:

- 2001-2002 The F. E. Smith Award in Communications Engineering Theory, Queen's University.

- 2001-2002 OGSST (Ontario Graduate Scholarship in Science and Technology).

- 2000-2001 OGS (Ontario Graduate Scholarship).

- 1999-2000 OGSST (Ontario Graduate Scholarship in Science and Technology).

- 1994-1996 NSERC (Natural Sciences and Engineering Research Council of Canada) Graduate Student Fellowship.

- 1993 NSERC (Summer project).

PUBLICATIONS:

- JOURNAL PAPERS

  1. Z. Rached and M. Racine, "Jordan Triple Systems of Degree at Most 2," *Communications in Algebra*, 24(3), 963-1001 (1996).

  2. Z. Rached and M. Racine, "Exceptional Jordan Triple Systems," *Communications in Algebra*, 25(8), 2687-2702 (1997).

  3. Z. Rached, F. Alajaji and L. L. Campbell, "Rényi's Divergence and Entropy Rates for Finite Alphabet Markov Sources ," *IEEE Transactions on Information Theory*, Vol. 47, No. 4, 1553-1560, May 2001.

  4. F. Alajaji, P.-N. Chen and Z. Rached, "A Note on the Poor-Verdú Conjecture for the Channel Reliability Function," *IEEE Transactions on Information Theory*, Vol. 48, No. 1, 309-313, January 2002.

  5. Z. Rached, F. Alajaji and L. L. Campbell, "On the Kullback-Leibler Divergence Rate for Finite Alphabet Markov Sources," submitted.

  6. F. Alajaji, P.-N. Chen and Z. Rached, "Csiszár's Cutoff rates for the general hypothesis testing problem," Under preparation.

- CONFERENCE PAPERS

1. Z. Rached F. Alajaji and L. L. Campbell, "Rényi's entropy rate for discrete Markov sources," *Proc. Conference on Information Sciences and Systems*, Baltimore, MD, March 17–19, 1999.

2. Z. Rached, F. Alajaji and L. L. Campbell, "On the Rényi divergence rate for finite-alphabet Markov sources," *Proc. International Symposium on Information Theory and Its Applications*, Honolulu, Hawaii, November 5–8, 2000.

3. F. Alajaji, P.-N. Chen and Z. Rached, "On the Poor-Verdú Conjecture for the Reliability Function of Channels with Memory," *Proc. International Symposium on Information Theory*, June 24-29, Washington, DC, 2001.

4. Z. Rached, F. Alajaji and L. L. Campbell, "A formula of the divergence rate for a class of Markov sources," *Proc. Conference on Information Sciences and Systems*, Princeton, March 20–22, 2002.

5. F. Alajaji, P.-N. Chen and Z. Rached, "Csiszár's forward cutoff rate for testing between two arbitrary sources," *Proc. IEEE International Symposium on Information Theory*, Lausanne, Switzerland, June 29–July 5, 2002.

WORK EXPERIENCE:

- Instructor at Queen's University for Math 474-874: Information Theory for fourth year undergraduate students and graduate students (Fall 2001).

- Teaching Assistant at Queen's University (Fall 1997-Winter 2000).

- Instructor at Algonquin College (Ottawa): Calculus for second year engineering students (Fall 1997).

- Teaching Assistant at Ottawa University (Fall 1992-Winter 1996).